Psychological Bulletin

CONTENTS

Published Bimonthly by the
American Psychological Association

Vol. 54, No. 3

The Psychological Bulletin contains evaluative reviews of research literature and articles on research methodology in psychology. This JOURNAL *does not publish reports of original research or original theoretical articles.*

Manuscripts should be sent to Wayne Dennis, Department of Psychology, Brooklyn College, Brooklyn 10, New York.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the "Publication Manual of the American Psychological Association" (*Psychological Bulletin*, 1952, 49 [No. 4, Part 2], 389–449). Special attention should be given to the section on the preparation of the references (pp. 433–440), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references. All manuscripts should be submitted in duplicate.* Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail.

Reprints. Fifty free offprints are given to contributors of articles and notes. Authors of early publication articles receive no gratis offprints.

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D. C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Menasha, Wisconsin
and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second-class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage under the provisions of Sec. 34.40 Par. (E) provided for in Section 538, act of February 28, 1925, authorized August 6, 1947. Printed in U.S.A.

VOL. 54, NO. 3 MAY, 1957

Psychological Bulletin

CONSTRUCT VALIDITY AND THE TAYLOR ANXIETY SCALE

RICHARD JESSOR AND KENNETH R. HAMMOND
University of Colorado

Construct validity (**9**, **31**) is an important new concept which has immediate implications for both psychometrician and experimentalist. Most important is the increased emphasis which construct validity places upon the role of theory in the validation of psychological tests. The aims of the present paper are two: (*a*) to consider the directive role of theory in the construction of psychological tests; and (*b*) to examine certain methodological issues which arise from the more explicit use of theory in test construction. For illustrative purposes we have chosen to make a critical analysis of the Taylor Anxiety Scale (*A* scale) and the research (**14**, **26**, **27**, **28**, **30**) in which it has been employed to establish the independent variable of drive (Hull's *D*).

The nature of the *A* scale and the results of the studies in which it has been used are well enough known so that only a brief description is necessary here. The scale is a self-report inventory consisting of 50 manifest-anxiety items and 175 buffer items, both groups of items taken almost entirely from the Minnesota Multiphasic Personality Inventory (MMPI). The research studies, concerned with testing the assumption that the *A* scale measures drive level, have evaluated the energizing property of *D*. They have indicated that, where the correct response in an experimental learning situation has a high probability of occurrence, the high scorers on the *A* scale perform better than the low scorers. Where the experimental situations are such that there are competing responses or the incorrect responses are equally likely of occurrence at the outset, the high scorers perform less adequately than the low scorers. Both findings are consistent with the Hullian assumption that all habit tendencies elicited in a given situation are multiplicatively affected by the level of drive at the time. These findings have provided the basis for inferring that the *A* scale is, therefore, a measure of drive.

THE USE OF THEORY IN TEST CONSTRUCTION

Cronbach and Meehl state that "Construct validation takes place when an investigator believes that his instrument reflects a particular construct to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses which are a means of confirming or disconfirming the claim" (**9**, p. 290). These authors take as the starting point of their discussion the presence of an already existing test or scale purporting to measure or thought to measure a particular variable. They are concerned with the methods of establishing the construct validity of a test after the test has been devised. The present authors, on the other hand, are concerned with the process of devising a scale or test so that it will be consistent with the procedures of construct validation. Our contention is that the test situation itself, and the kinds of test behavior it elicits, must be coordinated to the theory in exactly the same manner as the experiments aimed at validating the test. The Iowa experiments with the *A* scale were designed to fit the paradigm required by the Hullian framework—i.e., they were designed to measure learning, to control probability of occurrence of correct responses, to control other significant sources of drive variation, etc.—in order to make inferences to that framework. The same logic requires that the *A* scale itself should likewise have been designed so that performance on it might be a basis for inferring drive independently of the outcome of subsequent experiments.

Emphasis on the need for theoretical derivation of psychological tests may be found in recent work by Peak (**22**) and Butler (**5**). Their general contention is that the theory or the properties of the construct should determine the nature of the test itself as well as the nature of experiments which establish the construct validity of the test. Peak asserts that "The design of objective instruments and procedures requires . . . a theory about the characteristics and relationships of any variable to be measured . . . " (**22**, p. 296). She offers an enlightening example of this point: "If, for example, [the investigator] sets out to devise a measure of hostility with a knowledge of the psychoanalytic theory of defense mechanisms, the questions asked and the behavior observed will be very different from that which would seem relevant if manifest expressions of hostility were regarded as the only appropriate data" (**22**, p. 247).

Butler (**5**) has recently called attention to the preoccupation of psychometricians with the formal requirements of testing at the cost of ignoring the role of psychological theory in developing tests. He finds it astonishing that there is " . . . no personality inventory for which the content, the form of the items, and the psychometric methods applied have been dictated by a formal psychological model" (**5**, p. 77). The remainder of his article is a programmatic effort to use Tolman's theoretical model as the source of hypotheses about the nature of psychometric items most likely to provide useful intervening or independent variables.[1]

The important point here, which relates this discussion to construct validity, is that the psychological, or theoretical, model has implications for the psychometric, as well as for the experimental, procedure. It is an artifact of tradition that theories have been utilized to derive experiments but not to derive tests. Yet construct validity makes the same set of demands on both the psychometric and experimental approaches. Each approach requires that behavior take place under specified and controlled conditions. There seems to be no fundamental reason why theories should make unequivocal demands on the experiment and permit the test to satisfy psychometric requirements only.

The difficulties in moving from theories to empirical conditions, or from theories to classes of observable

[1] Although it is not the purpose of this article to examine examples of tests whose items have been derived from theoretical models, the reader may refer to one such test which will serve as an illustration. The test was derived by Liverant (**20**) from Rotter's social learning theory (**24**) in order to measure the construct of need value.

behaviors, are, of course, apparent. Peak (**22**) acknowledges that there is no simple methodological prescription for meeting the requirements of theories. Cronbach and Meehl (**9**) call attention to the absence in psychology of a formal calculus which can provide rigorous implicit definitions of primitive terms and give them empirical meaning. Nevertheless, as they point out (**9**, p. 294), a theoretical network, though admittedly vague and sketchy, does exist and provides constructs with whatever meaning they do have.

This network, which guides attempts at construct validity, should play also the *prior* role, we suggest, of guiding test development. Such procedure would have important implications for the adoption of strategy in subsequent construct-validation attempts where the outcome proves to be negative and the investigator has to decide where to lay blame—on the test or the theory—and decide which to revise or discard.

The Development of the Taylor Anxiety Scale

With the foregoing considerations in mind, we return to an examination of the development of the *A* scale. The general question we are asking about the *A* scale is: In what way are the form of the scale, the item selection procedure, the item content, and the nature of the responses elicited by the scale coordinated to or derived from the Hullian framework as indicants of drive. Nowhere, to our knowledge, is this made explicit or is a suitable answer to be found; yet this is precisely what our point of view would demand.

Form of the Scale

The issue in this section lies in the coordination between the inventory self-report form of the *A* scale and the Hullian construct of drive. In Hullian theory, drive level is coordinated to both antecedents and consequents. The antecedents are generally conditions, e.g., food deprivation, shock, etc., which establish internal states that the organism seeks to avoid. The consequents of drive level are activity or level of energy expenditure. It is clear that the inference of drive level from the *A* scale is contingent upon consequents; i.e., drive level in this case is a response-inferred construct, since no control or manipulation of conditions antecedent to the *A*-scale responses has been accomplished. Although there has been some general criticism of response-inferred constructs (**18**, **25**), it is clear that Brown and Farber (**4**) do not consider such criticism fully warranted with respect to inferring drive. They note (according to Farber) that while " . . . more data than those provided by the topography of a response are needed to enable one to identify the extent of its dependence upon one rather than another of its many determinants, this does not mean that there are *no* criteria of drive applicable to responses" (**12**, p. 26). This is an important statement; yet the obvious fact is that such criteria are nowhere presented in a manner which would coordinate inventory self-reports to drive. Their statement suggests the future possibility of reliance on inventories, but a query still has to be raised as to whether the Hullian concept of drive can, in terms of its present definition, be at all coordinated to self-report verbal responses on any inventory.

Selection of Items

If one is concerned only with the predictive validity of a test, the matter of item content is relatively unimportant, for the empirical item-criterion correlations provide criteria for the final selection of items. However, when a test-developer insists (cf. **28**, p. 84; **13**, p. 324) that his purpose includes more than the prediction of a particular criterion performance and that the test items are intended to be indicators of a construct, then item content becomes highly important, and item-criterion correlations only are insufficient.

No one can say precisely what the specific steps relating empirical operations to a construct should be, since these must vary with the nature of the construct and the intent of the investigator. However, it is possible to assert that the chain of empirical operations should meet at least one criterion. This criterion is made explicit by Cronbach and Meehl as follows: " . . . unless the network . . . [of constructs and hypotheses] exhibits explicit, public steps of inference, construct validation cannot be claimed" (**9**, p. 291). We take this to mean that all the methodological links in the development of a test must be scrutinized for their "explicit, public," and therefore objective and retraceable, character. No test can be more objective than the most subjective link in its development.

Therefore, test items which are intended to indicate a construct should be selected by rational (rather than intuitive) means. This means that an item should be scrutinized for its logical relationship to a construct and that the grounds for choice of an item should be explicit and public. The difficulty of deriving a series of items will depend on the scope, precision, content, etc., of the construct to be measured; but there should be no need to resort to a procedure which relies on implicit and private (i.e., undefended or unexplained) judgments or ratings. A single explicit (and sound) argument for an item is better than an implicit rating of an item by many judges, because the former is a retraceable (and thereby self-corrective) step while the latter is not. (Once the choice of items has been made, the empirical criterion for inclusion may well be interitem correlations since the concept may specify a unitary function.) In any case, high interobserver agreement is no substitute for logical validity.

This point needs emphasis because clinical psychologists and psychiatrists are often used as judges in the development of tests. Since their judgments are usually obtained on an intuitive basis (the judges are rarely asked to deduce the items according to the logical requirements of a concept), a hazard is created (in reference to construct validity) which cannot be overcome by appeal to authority (cf. **16**).

The hazard introduced by the intuitive procedures usually involved in judging is exemplified in the lack of explicit relationship between *A*-scale items and drive properties. As pointed out above, Taylor's concept of drive was intended to be identical with Hull's. Yet the procedure for selecting items apparently was *not* to scrutinize them for their logical relationship to drive. Rather, the items were selected on the basis of clinical impression of how well they fit Cameron's (**7**) definition of anxiety. But why Cameron's definition of a concept when it is Hull's concept of drive which is to be given empirical content? An examination of Cameron's definition leads us to conclude that there are no obvious reasons for choosing his definition rather than any other.

Subsequently the test was shown to discriminate to some extent between psychiatric patients and normals.

Taylor reports that, "In an attempt to determine the relationship between the anxiety-scale scores and manifest anxiety as defined and observed by the clinician, the anxiety scores for groups of normal individuals and psychiatric patients were compared" (**29**, p. 290). The empirical situation at this point is as follows: The scale can now be said to be representative of certain clinicians' judgments about patients, i.e., the scale is a quick device for reaching the same decision as certain clinicians about the manifest anxiety of patients. Unfortunately, it is by no means clear what relationship this classification by clinicians has to the Hullian concept of drive.

Content of the Items

One might well question how the content of various items of the *A* scale can be conceptualized in terms of the properties of drive. Why should answering "false" to such statements as "I have very few headaches" and "I am very confident of myself" constitute a referent for higher drive than answering them "true"? Why should answering "true" to an item reporting diarrhea and one reporting constipation ***both*** indicate higher drive than answering "false" or answering one of them "true" and the other "false"? Answers to our questions about the content of the items would be provided whenever a test has been derived from a theory. Taylor states that two assumptions guided the use of the *A* scale: "First, that variation in drive level of the individual is related to the level of internal anxiety or emotionality, and second, that the intensity of this anxiety could be ascertained by a paper-and-pencil test consisting of items describing what have been called overt or manifest symptoms of this state" (**29**, p. 285). The first assumption is pertinent here. Only if one is willing to equate emotionality or internal anxiety with level of energy expenditure could one accept the use of some of the items. This equation itself requires logical justification. The actual procedure, however, was to have judges select items relating to manifest anxiety, and a logical gap thus exists between manifest anxiety and energy expenditure.

Nature of the Responses

The second assumption Taylor mentions is that anxiety or emotionality may be assessed by a paper-and-pencil test in which the subject acknowledges symptoms of this state. Under our discussion of the form of the scale we raised questions about the theoretical soundness of this assumption. Here we wish to turn our attention to psychometric aspects of this assumption. In experimental situations we generally observe or measure the actual behavior or responses on which we base our inferences. The same is not true in psychometric measurement. We observe or elicit responses (verbal) about other responses (nonverbal). A veridical relationship between verbal and nonverbal responses is a fundamental requirement of the chain of inference involved in the use of the *A* scale to measure drive. Yet the *A* scale is vulnerable to the oft-cited and well-substantiated criticisms of self-report inventories where the social desirability or meaning of the item content is clear to the respondent. Procedures designed to maximize veridicality, such as forced-choice items, or to detect lack of honesty, such as the L and K scales on the MMPI, are not utilized in the test. (Since the items from the latter scales are included among the buffer items, their use is apparently left to the discretion of the test user.) As a matter of fact, a recent study (**10**) suggests that the Taylor scale may be more susceptible to deception than are other objective measures for measuring anxiety.

Further, the *A* scale elicits only two responses: "true" or "false." Since the purpose of the scale is to arrive at a measure of intensity of anxiety, a scale form providing for responses of varying intensity for each item would seem preferable. A rating scale for each item, or at least several response categories ranging in degree of agreement or disagreement with the item, might be more appropriate.

Thus far we have examined the development of the *A* scale in terms of its relationship to Hullian theory. The lack of any logical relationship raises the issue of interpretation of research findings. One might ask what the consequences would have been, for either theory or test, had the studies with the *A* scale yielded negative findings. Cronbach and Meehl (**9**, p. 295) note that the investigator whose prediction and data are discordant must make strategic decisions: he may decide his test is not an adequate measure of the construct, or he may call into question the network defining the construct, if he has confidence in the test. The latter phrase is the core of the matter. With respect to the *A* scale, no explicit logical basis for confidence in the test as a measure of drive existed prior to the experiments, and the strategy of the investigator, in the face of negative results, would undoubtedly have been to challenge the test rather than Hull's assumptions. The issue, however, would be more critical where both the test and the nomological network are at their inception and neither has been extensively employed, because the reason for negative findings is equally likely to be in the theory or the test. In such cases (typical for personality formulations) a theoretically derived test would yield the advantage of directing further theoretical analysis and development.

The Construct Validity of the *A* Scale

We shall next consider the status of the *A*-scale studies within the framework of construct validity as discussed by Cronbach and Meehl. This section will be limited to two main points: (*a*) the degree to which diverse aspects or consequences of the nomological net surrounding the construct of drive have been investigated; and (*b*) the degree to which alternative inferences from the *A*-scale studies have been disconfirmed.

Validation of Diverse Properties of a Construct

Cronbach and Meehl stress the following point: "Numerous successful predictions dealing with phenotypicallydiverse 'criteria' give greater weight to the claim of construct validity than do fewer predictions, or predictions involving very similar behaviors. In arriving at diverse predictions, the hypothesis of test validity is connected each time to a subnetwork largely independent of the portion previously used. Success of these derivations testifies to the inductive power of the test-validity statement, and renders it unlikely that an equally effective alternative can be offered" (**9**, p. 295). And, "The test developer must investigate far-separated, independent sections of the network" (**9**, p. 299). Further, in a related discussion of the establishment of connections between inferred entities and observables, Beck emphasizes the methodological rule that "Each component of the inferred entity must be symptomized by some datum, actual or available . . . " (**2** p. 375).

In a recent paper Farber acknowledges two components or properties of drive—energizing and reinforcing. He indicates that a given variable has the characteristics of a drive "if a) its elimination or reduction in magnitude is reinforcing, and/or b) it has a general dynamogenic effect upon the response tendencies elicited in a given situation" (**12**, pp. 38–39). The bulk of animal studies have dealt with the former property. None of the *A*-scale studies has investigated this property—all have, instead, dealt with the latter, energizing, property of drive (the multiplicative relation of $_{S}H_{R}$ and *D*). Farber notes that although there are difficulties in demonstrating the reinforcing properties of manifest anxiety, "It is quite possible that this sort of demonstration can be accomplished, but to the best of my knowledge no one has yet done so" (**12**, p. 27). The requirements of construct validation would certainly favor the exploration of this "far-separated, independent section of the network."

Other sections of the "phenotypic space" also require investigation, particularly the effect of manipulation of antecedent conditions on the *A*-scale responses themselves. Atkinson (**1**) asks whether scores on the *A* scale would increase if anxiety were experimentally increased. The possibility of employing conditions, e.g., shock, suggested by other research concerned with establishing or increasing drive, becomes apparent. To summarize, the point is that construct validity requires investigation of diverse properties of the construct. One reason for making this requirement is to lower the likelihood of finding acceptable alternative inferences which can encompass such diversity. This leads us to the next major issue.

Disconfirmation of Alternative Inferences

Confirmation of an inference is also established to the extent that other inferences are not equally applicable. Beck states that "Confirmation can come only from the disconfirmation of all alternative hypotheses through the evidential denial of at least one consequent of each alternative . . . " (**2**, p. 377). In the light of this criterion, the inference of drive from the *A*-scale studies is not secure. Various investigators have made alternative inferences about what the Taylor scale measures. Three of these alternatives will be mentioned here.

(*a*) Most prominent is the controversy raised by Hilgard (**17**), and by Child (**8**). They consider an equally plausible hypothesis to be that the *A* scale measures only different $_{S}H_{R}$'s rather than different drive levels. Hilgard has concluded that anxiety responses or anxiety-related responses, e.g., stronger defensive or avoidance habits, can account equally well for the data. Certainly, on the face of it, the scale measures nothing other than differential response systems (assuming veridicality). Farber has been explicit in acknowledging an associative component in what the *A* scale measures, but he insists that it is the drive component which is inferable from the research. Overlooking the possibility, as suggested by Postman (**23**), that there are no operational means for separating these two components, our immediate purpose is to indicate that alternatives to the drive inference have, at the very least, as yet not been disconfirmed.

(*b*) Recent studies (**6**, **15**, **19**) have also suggested another alternative hypothesis, namely that the scale measures intellectual (habit?) differences rather than drive. While the implications of intelligence as an explanation for the *A*-scale findings are not yet clear, these empirical findings should be considered. Certainly the obtained correlations between *A*-scale scores and intelligence (if they are not fortuitous) are not referable to any property of drive as thus far defined.

(*c*) Finally, the near-perfect correlation between the *A* scale and the MMPI psychasthenia scale (**3**, **11**), and correlations with other neurotic inventories (**10**), raises further questions as to whether *any* neurotic inventory would yield similar experimental findings, and, if so, in what way neuroticism in general is coordinated to drive level.

For the test to meet fully the requirements of construct validity, these alternative inferences must be disconfirmed.

The Conditional Definition of Anxiety as Drive

To illustrate a further methodological point concerning construct validity, let us assume our argument concerning the *A* scale to be valid: that when it was used in connection with the test of the Hullian hypothesis that drive energizes $_{S}H_{R}$'s, the *A* scale had neither logical nor empirical foundation as a measure of drive. This raises the question of whether the studies achieved a definition of drive or tested the Hullian hypothesis.

The first step in the experiment was to identify high and low scorers on the *A* scale. (Note that no theoretically relevant meaning can be given to any score at this point because no logical or empirical tie can be made to the Hullian concept of drive.) The next step was to perform the experiment in the eyelid conditioning arrangement. Results conforming to expectations were obtained. The researchers had then obtained a ***conditional definition*** of the concept of drive. The definition is of this form: if a high scorer is placed in a conditioning arrangement, then a high scorer has a high drive if, and only if, the high scorer conditions more rapidly than a low scorer conditions. Farber puts this as follows: " . . . the question of the validity of the Taylor scale *as a useful definition of general drive level* is answered by the accuracy of the prediction of relations between this scale and specified behavior variables, under conditions such that variation in the behavioral variables can be reasonably attributed to differential drive levels" (**13**, p. 325). Thus, it appears that the researchers were attempting to achieve a conditional definition of drive.

The investigators assert, however, that the experimental situation provided both a definition of drive and a test of the Hullian hypothesis under consideration at one and the same time. This procedure introduces an ambiguity. If the results are negative, does it mean that the definition of the concept is faulty, or the hypothesis relating the variables? A difficulty remains if the results are positive, since the hypothesis is proved only by asserting the definition. That is, if we ask how the experimenter knows that he actually varied drive, he can only reply that the results are meaningful *if* the *A* scale measures drive. However, as we have pointed out, the construction of the *A* scale has not been coordinated to drive, the scale has not been employed in testing the diverse properties of drive, nor have "reasonable" alternative inferences about the *A*-

scale scores been disconfirmed. Therefore the experimental results alone are not sufficient support for the assumption that the *A* scale measures drive.

The problem of the definition of drive is directly analogous to the problem of the definition of reinforcement. Meehl (**21**) points out that the reason why the Law of Effect is not circular is that conditional definitions of reinforcers are made independently of the test of the Law of Effect. For example, Meehl presents the following "special law": "On schedule M, the termination of response sequence R, in setting S, by stimulus S^1 is followed by an increment in the strength of S.R." (**21**, p. 60). In the studies involving the *A* scale, however, M (the *A* scale) is not defined by the investigators in terms of an independently observed "schedule," but only in terms of the response sequence in the experiment.

When a construct implies a relationship between variables, these variables must be designated independently of any test of that relationship.

Summary

Construct validity emphasizes the directive role of theory in test validation; the intent of this paper has been to emphasize the directive role of theory in the construction of psychological tests.

Our position is that the psychometric, as well as the experimental, procedure must be coordinated to the hypothetical properties of the construct to be measured. In this way the test situation is made parallel to the experimental situation—the conditions of both being clearly derived from theory, and the behavior elicited in both being clearly relevant to the theory.

The above points, as well as certain methodological issues arising from the explicit use of theory in test construction—the investigation of diverse properties of a construct, disconfirmation of alternate inferences, conditional definitions—were illustrated through a critical examination of the Taylor Anxiety Scale. Our conclusion was that the *A* scale has only a tenuous theoretical and empirical coordination to the Hullian construct of drive. The experiments which have relied on the *A* scale may be considered to have attempted thus far a conditional definition of drive rather than to have demonstrated the hypothesis that drive energizes habits.

Our intent has not been to single out a particular test for criticism. We recognize that much work is being done on further construct validation of the *A* scale. Such work, we hope, will answer some of the questions we have raised, questions which we feel are of importance for psychometrics as a whole.

REFERENCES

1. Atkinson, J. W. Comments on Professor Farber's paper. In *The Nebraska symposium on motivation*. Lincoln, Nebr., Nebraska Univer. Press, 1954. Pp. 51–55.
2. Beck, L. W. Constructions and inferred entities. In H. Feigl & M. Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 368–381.
3. Brackbill, G., & Little, K. B. MMPI correlates of the Taylor Scale of Manifest Anxiety. *J. consult. Psychol.*, 1954, **18**, 433–436.
4. Brown, J. S., & Farber, I. E. Emotions conceptualized as intervening variables—with suggestions toward a theory of frustration. *Psychol. Bull.*, 1951, **48**, 465–480.
5. Butler, J. M. The use of a psychological

model in personality testing. *Educ. psychol. measmt*, 1954, **14**, 77–89.

6. CALVIN, A. D., ET AL. A further investigation of the relationship between manifest anxiety and intelligence. *J. consult. Psychol.*, 1955, **19**, 280–282.

7. CAMERON, N. *The psychology of behavior disorders.* New York: Houghton Mifflin, 1947.

8. CHILD, I. L. Personality. *Ann. Rev. Psychol.*, 1954, **5**, 149–171.

9. CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

10. DAVIDS, A. Relations among several objective measures of anxiety under different conditions of motivation. *J. consult, Psychol.*, 1955, **19**, 275–279.

11. ERIKSEN, C. W., & DAVIDS, A. The meaning and clinical validity of the Taylor Anxiety Scale and the hysteria-psychasthenia scales from the MMPI. *J. abnorm. soc. Psychol.*, 1955, **50**, 135–137.

12. FARBER, I. E. Anxiety as a drive state. In *The Nebraska symposium on motivation.* Lincoln, Nebr.: Nebraska Univer. Press, 1954. Pp. 1–46.

13. FARBER, I. E. The role of motivation in verbal learning and performance. *Psychol. Bull.*, 1955, **52**, 311–327.

14. FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 120–125.

15. GRICE, G. R. Discrimination reaction time as a function of anxiety and intelligence. *J. abnorm. soc. Psychol.*, 1955, **50**, 71–74.

16. HAMMOND, K. R. Probabilistic functioning and the clinical method. *Psychol. Rev.*, 1955, **62**, 255–262.

17. HILGARD, E. R. Theories of human learning and problems of training. In *Symposium on psychology of learning basic to military training problems.* Panel on Training and Training Devices. Res. & Develpm. Bd, 1953.

18. JESSOR, R. Phenomenological personality theories and the data language of psychology. *Psychol. Rev.*, 1956, **63**, 173–180.

19. KERRICK, JEAN S. Some correlates of the Taylor Manifest Anxiety Scale. *J. abnorm. soc. Psychol.*, 1955, **50**, 75–77.

20. LIVERANT, S. The use of Rotter's social learning theory in the development of a personality inventory. Unpublished doctor's dissertation, Univer. of Colorado, 1956.

21. MEEHL, P. E. On the circularity of the law of effect. *Psychol. Bull.*, 1950, **47**, 52–75.

22. PEAK, HELEN. Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden Press, 1953. Pp. 243–300.

23. POSTMAN, L. J. Comments on papers by Professors Brown and Harlow. In *Current theory and research in motivation.* Lincoln, Nebr.: Nebraska Univer. Press, 1953. Pp. 55–58.

24. ROTTER, J. B. *Social learning and clinical psychology.* New York: Prentice-Hall, 1954.

25. SPENCE, K. W. The nature of theory construction in contemporary psychology. *Psychol. Rev.*, 1944, **51**, 47–68.

26. SPENCE, K. W., & TAYLOR, JANET A. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *J. exp. Psychol.*, 1951, **42**, 183–188.

27. SPENCE, K. W., & FARBER, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 116–119.

28. TAYLOR, JANET A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, **41**, 81–92.

29. TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

30. TAYLOR, JANET A., & SPENCE, K. W. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, **44**, 61–64.

31. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Suppl.*, 1954, **51**, 2, Part 2, 1–38.

Received March 5, 1956.

Psychological Bulletin
Vol. 54, No. 3, 1957

A BIOMETRIC EVALUATION OF THE SOMATOTHERAPIES IN SCHIZOPHRENIA

VIRGINIA M. STAUDT[1]
Hunter College

and JOSEPH ZUBIN
Biometrics Research, State of New York Department of Mental Hygiene, and Columbia University

Although the need for sound research on the outcome of mental disease has long been recognized, and although probably no other medical specialty has accumulated such an abundance of statistics as psychiatry, yet the dearth of adequately designed investigations especially in the field of the evaluation of therapy is conspicuous. The literature describing and appraising the results of insulin, metrazol, and electroshock treatment is voluminous. Reports on lobotomy results have also mounted steadily. In spite of these studies the satisfactory evaluation of therapy still remains one of the most troublesome problems of psychiatric research. In reviewing the vast literature of therapeutic results one finds conflicting reports ranging from severe skepticism of the various therapies to inordinate enthusiasm for them. Consequently every clinician can cite a study to support his particular viewpoint on any given therapy. This state of affairs suggests the need for a critical study of the methodology of these evaluation studies with a view towards the improvement of research design in this area in the future.

The Problem

It would be a Herculean task to attempt to record and to review all the studies that have been done on the evaluation of the various somatotherapies, not only because of their excessive number but also because their results, so diversely presented, defy organized classification according to any one uniform plan. It is, therefore, our intention to present a fairly representative group of researches with their results and the techniques by which these have been derived.

The purpose of this study is to examine the available and analyzable data on the outcome of the shock therapies and psychosurgery in order to get an estimate of the effectiveness of these therapies in the treatment of schizophrenia, as it has been reported in the literature. As a result, four types of data have been analyzed: (*a*) outcome of nonspecific treatment of all psychoses and schizophrenia during the preshock (pre-1930) period; (*b*) outcome of nonspecific treatment of schizophrenia during the preshock period, as reported after shock therapy became available; (*c*) outcome of insulin, electroshock, metrazol, and psychosurgical therapies; and (*d*) results of comparative studies of treated and control groups.

Outcome of Nonspecific Treatment of All Psychoses and Schizophrenia During the Preshock (Pre-1930) Period

In the beginning of the 20th century both in America and abroad,

[1] The present publication has resulted from a study undertaken by this author during her tenure of a fellowship granted by The Fund for the Advancement of Education. The facilities of Project M586 of the National Institute of Mental Health aided in the preparation of the material on which this study is based.

interest was directed toward evaluating the outcome of mental disease. Practical considerations, such as the cost of care of the mentally ill to the community, as well as scientific curiosity prompted hospital administrators and doctors to investigate the results of hospitalizing and caring for mental patients. Most studies consisted in the follow-up of total hospital populations over a period of several years after admission to determine the ultimate disposition of these cases. Bond, Fuller, and Pollock were pioneers in this research. All three studied the outcome of mental disease and all three were unanimous in finding a low rate of improvement in dementia praecox patients.[2] We shall review the work of Bond and Fuller in considerable detail.

In several studies Bond (**6**, **7**, **8**, **9**) reported his findings on the then current results for use as a base line or standard by which new therapeutic measures could be judged as they were developed. Recognizing the scarcity of follow-up results in psychiatric work as compared with surgery, he compiled the data which he had collected mainly at the Pennsylvania Hospital, a private institution. Bond observed that usually the patients were committed to the Hospital after many opportunities for early treatment had been lost because of the families' tendencies to procrastinate. The summary of Bond's early follow-up studies on heterogeneous groups of patients in the preshock period is presented in Table 1.

In general Bond considered his findings encouraging and observed that, although the cases came late for treatment, if the psychiatrist could still produce recovery in approximately 25% with about at least 15% ameliorated, then he might rightfully feel gratified.

From Table 1 it can be seen that the recovery and improvement rate combined is between 40% and 50%. Bond maintained that these good results with mental patients should be emphasized to counteract the then (1920's) popular impression that mental patients always become worse and that even if they seem to recover they soon break down again. A point that one cannot fail to note here in reviewing Bond's results is the heterogeneity of the patients. All ages, all

TABLE 1

BOND'S FINDINGS ON THE OUTCOME OF MENTAL DISEASE IN PATIENTS HOSPITALIZED DURING THE PRESHOCK PERIOD*

Year	Study	Patient		Duration of Follow-up	Results (%)		
		Type	N	(P.Ad.)	R & I	U	D
1921	Bond (**6**)	V^1	111	5 years	49	30	21
1921	Bond (**7**)	V^2	251	5 years	54	25	21
1923	Bond (**8**)	V^3	377	5 years	43	30	27
1925	Bond (**9**)	V^4	1024	5 years	41.2	25.5	32.3

Note.—V^1 includes DP, MD, SA, GP, U, and C patients.
V^2 includes DP, MD, IM, SA, D, G, OBD, E, A, P.som., Pa, PN, Mo-h, P.inf., PsPath.
V^3 includes SA, GP, MD, DP, OBD, and P.pell. patients.
V^4 includes MD, DP, GP, SA, IM, PN, Pa, PsPath, P.som. patients.

* For glossary of abbreviations used in this and in succeeding tables see Key to Tables in Appendix.

[2] It would be interesting to determine whether the uniformly low rate of improvement in these early studies was a consequence of the more narrow definition of dementia praecox which Kraepelin introduced, rather than the wider definition introduced later by Bleuler under the term "schizophrenias."

types of diseases, and individuals with illnesses of varying durations are represented here, but these are followed up for a long period. While Bond was interested primarily in the general outcome for all mental diseases, he mentions specifically that in his groups the most unchanged by treatment were the dementia praecox cases. The above tabulated studies are important for they were later used in several studies as base lines in evaluating some of the results of the shock therapies in the 1930's.

In addition to Bond, Fuller also studied the outcome of mental illness. Among his many surveys during the preshock period, he reported in 1930 on the expectation of hospital life and outcome for mental patients for first admissions to mental hospitals (**24**). This extensive survey furnished data on a wide variety of psychoses as well as a report on all psychoses. He included 1,200 patients in each group of psychoses such as Dementia Praecox or Manic Depressive, and 2,400 in the "All Psychoses" group. Fuller presented his statistics in a particularly effective way, giving the results in terms of patients discharged, dead, and still hospitalized at various time intervals up to 15 years. Fuller's findings regarding the outcome of first admissions to a mental hospital are presented in Table 2.

In Fig. 1 the percentage discharged and not later readmitted for All Psychoses and for Schizophrenia is presented. It will be noted that from 3 months to 15 years there is a rather steady rise in the percentage discharged and not later readmitted for both groups, the rate being slightly higher for All Psychoses than for Schizophrenia. At the end of 5 years, 29.9% in Schizophrenia and 36.9% in the All Psychoses categories respectively are discharged and not later readmitted. At the end of 10 years, 32.2% of the Schizophrenic group and 39.3% of All Psychoses are discharged and not later readmitted, while after 15 years the percentages are 35.3 and 40.9 respectively for Schizophrenia and All Psychoses.

In another survey (**25**) based on 11,050 patients admitted over a two-year period to the civil State hospitals of New York, Fuller found that out of every 1,000 patients representing first admissions of all diagnostic categories 87.1% were hospitalized only once, while 12.9% had more than one hospitalization. In 1931 in studying the duration of hospital life for mental patients, Fuller (**26**) reported the

TABLE 2

FULLER'S FINDINGS ON EXPECTATION OF HOSPITAL LIFE AND OUTCOME OF MENTAL PATIENTS ON FIRST ADMISSION

Follow-up Period	MD Percentages Based on $N = 1,200$				DP Percentages Based on $N = 1,200$				AP Percentages Based on $N = 2,400$			
	Not Later Readmitted	Later Readmitted	In	Dead	Not Later Readmitted	Later Readmitted	In	Dead	Not Later Readmitted	Later Readmitted	In	Dead
3 mos.	15.6	3.2	76.5	4.7	10.2	1.8	86.3	1.7	11.8	1.4	75.4	11.4
6 mos.	25.3	9.1	59.5	6.1	13.8	4.7	79.0	2.5	18.3	3.8	62.9	15.0
9 mos.	34.1	14.2	44.9	6.8	16.4	6.8	73.3	3.5	23.4	5.7	53.3	17.6
1 yr.	41.1	17.5	34.4	7.0	18.5	8.3	69.4	3.8	27.1	6.5	46.5	19.9
2 yrs.	50.6	17.7	23.3	8.4	24.2	8.6	62.3	4.9	32.3	6.0	36.5	25.2
3 yrs.	54.7	18.0	17.8	9.5	27.0	7.2	59.8	6.0	35.1	5.2	31.3	28.4
4 yrs.	56.0	16.7	16.8	10.5	28.3	7.4	56.6	7.7	36.0	7.4	25.8	30.8
5 yrs.	57.1	16.0	15.6	11.3	29.9	7.2	54.0	8.9	36.9	4.7	25.8	32.6
10 yrs.	62.2	6.7	15.8	15.3	32.2	4.0	46.0	17.8	39.3	2.3	20.0	38.4
15 yrs.	66.8	2.4	13.1	17.7	35.3	1.3	38.4	25.0	40.9	.7	17.0	41.4

outcome for three groups of mental patients: one group admitted to New York State civil hospitals between 1909 and 1911 and observed for 16 years; another admitted from 1914 to 1916 and followed for 11 years; and the last admitted from 1919 to 1921 and followed for 6 years. In Table 3 Fuller's findings for these groups are presented. He gave the results for the individual psychoses and then presented the outcome for all psychoses as a group. Here as in the previous table we present his findings for manic depressive, dementia praecox, and the total of all psychoses.

Just prior to the introduction of shock therapy in the United States in 1935, Fuller (**27**) published another report on the outcome of mental disease in 947 patients discharged from the civil State hospitals of New York during the decade following their discharge. As a result of his investigation he was able to estimate that during a ten-year period, out of each 100 patients discharged from the civil State hospitals of New York, 55 would be living in the community, 21 would be resident again in a mental hospital and 23 would have died either in the community or in a mental hospital and 1 out of the total 100 would be located in some type of institution other than a mental hospital. A summary of his findings is presented in Table 4.

Once again, as in the case of the Bond studies, the groups used by Fuller were heterogeneous. Bond's recovered and improved category in Table 1 is quite similar to Fuller's percentage in the community after

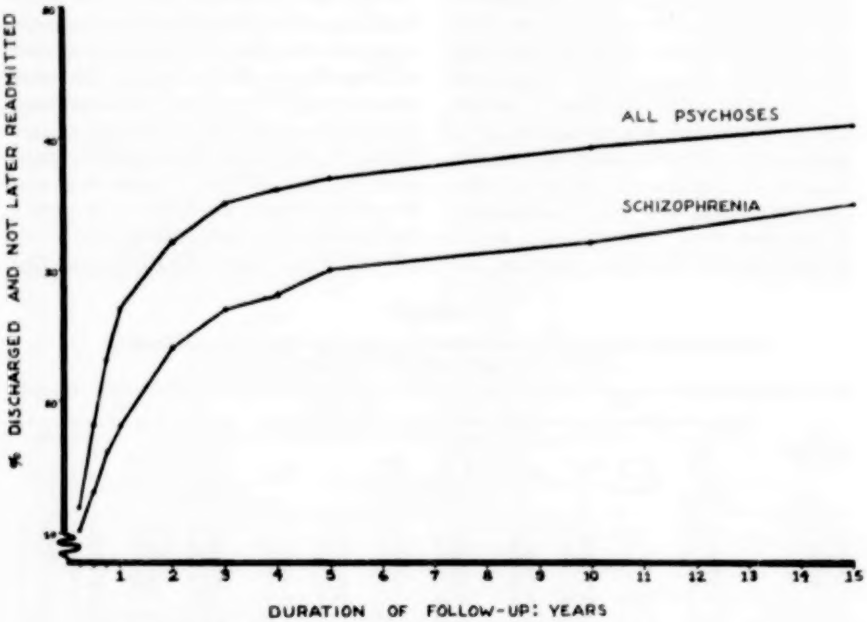

FIG. 1. FULLER'S FINDINGS ON OUTCOME OF MENTAL PATIENTS ON FIRST ADMISSION (1930)

TABLE 3

Fuller's Findings (1931) on the Duration of Hospital Life for Mental Patients

Type	*N*	Duration of Follow-up (P. F. Ad.)	Results (%)			In Hosp. at End of Period
			R & I	U	D (In Hosp.)	
MD	1,579	16 years	62.0	6.2	18.4	13.4
DP	2,481		25.0	12.7	26.5	35.8
AP	11,050		34.7	7.4	42.1	15.8
MD	1,868	11 years	65.0	2.3	17.7	15.0
DP	3,549		28.0	9.3	21.3	40.8
AP	12,550		33.0	5.4	42.0	19.6
MD	1,873	6 years	61.8	3.5	17.6	17.1
DP	4,119		27.4	8.2	13.0	51.4
AP	13,473		31.5	5.3	38.1	25.1

ten years in Table 4. The unimproved and death rates also tend to be similar.

Outcome of Nonspecific Treatment of Schizophrenia During the Preshock Period, as Reported After Shock Therapy Became Available

In 1936 with the advent of the shock therapies to the United States, the need for comparative norms based on nonshock patients in the evaluation of the results of shock therapy became more acute. As psychiatrists working with the new techniques searched about, they found only a limited number of studies with which to compare the new results (except for those of Bond, Fuller, and Pollock), especially for homogeneous groups, as, for example, schizophrenics. Therefore, in the late 1930's several studies evaluating results during the preshock period were instituted. Table 5 presents the results of a group of such studies on the outcome of mental disease in patients, mainly schizophrenics, hospitalized during the preshock period. In appraising these results it must be realized that during that era patients received mainly routine hospital care, or what is frequently referred to now as non-

TABLE 4

Fuller's Findings (1935) on Outcome of 947 Patients Discharged from the Civil State Hospitals of New York Ten Years After Discharge

Type	*N*	In Community After Ten Years (P.Dis.)	In Hospital	Died During Period	In Community Continuously	Readmitted to a Mental Hospital
MD	327	56.6	19.9	22.6	37.3	47.7
DP	242	43.8	43.3	12.8	36.4	55.8
AOP	378	59.5	8.7	31.0	48.1	31.7
AP	947	54.5	21.4	23.4	41.6	43.4

Note.—AOP includes CA, Gpl, CS, OBND, A, D, P.som., MD, IM, DP, E, Pa, PN, N, PsPath, Mdef, Ud; AP includes all psychoses.

specific treatment. Thus, the recoveries that occurred under such conditions are called, according to present standards, "spontaneous remissions." Some psychiatrists, however, like Malamud (**52**) and Solomon maintain that there is no such thing as "spontaneous remission." They contend that every patient gets something out of his hospital stay and that something is what helps in recovery. Therefore, they claim that everything that is done for the patient is in one sense or another treatment, whether or not it is meant to be treatment by the doctor who administers it.

In opposition to this viewpoint, there are those investigators like Stunkard (**76**) who maintain that in order to evaluate therapeutic effects of a specific therapy it is necessary to contrast it with the expected spontaneous improvement. To prove the effectiveness of a particular therapy one should be able to demonstrate, other factors remaining constant, that the patient makes more progress with the therapy than without it. The essential problem in this "spontaneous remission" controversy revolves around the significance of the term "treatment." These studies of spontaneous remission seem[2] to be measuring the same factors as the studies of the so-called nonspecific treatments. It should be mentioned here that ordinarily nonspecific treatment includes hydrotherapy, recreational therapy, occupational therapy, physical therapy, and even brief interviews with physicians.

In Table 5 we have included the "spontaneous remission" and the nonspecific treatment studies. A review of these investigations reveals very little specification on the part of the authors of the type of patients, duration of illness, duration of follow-up and of other relevant factors necessary to compare different studies. The earliest study listed in the table, that of Bond and Braceland, reports the outcome based on a heterogeneous group. The later studies were based exclusively on patients suffering from schizophrenia, a disease so well represented in all mental hospitals that it offered the greatest challenge to the new therapies. Comparing the over-all results with the results for the schizophrenic group in the Bond and Braceland study, one finds a much higher recovery rate for the heterogeneous group. Improvement in the schizophrenic group is fairly low in all studies. Generally speaking, most of these studies indicate about a 30–40% improvement rate maintained upon follow-up by these so-called spontaneous remissions or by those who have had nonspecific treatment. Similar results have also been reported from abroad by Neumann and Finkenbrink (**60**) who found 32.9% social remissions after a twenty-year follow-up. The chief difficulty in appraising these results is that most workers have not indicated the duration of the psychosis at the time that the patients came for treatment. From the few studies in which the duration was indicated, it can be seen that there is considerable variability in this factor in the different studies, as in the case of Gelperin's (**28**) group which included both acute and chronic cases. Such variability makes comparisons very difficult. However, as we shall see presently, the shock period studies of the somatotherapies likewise included individuals who had been ill for varying lengths of time. The statistics that have been presented in the foregoing tables offer only an over-all view.

In summary then, it may be said in respect to the nonspecific treat-

TABLE 5

OUTCOME OF SCHIZOPHRENIA IN PATIENTS RECEIVING NONSPECIFIC TREATMENT, AS REPORTED DURING THE SHOCK PERIOD

Study	Patient		Duration of Follow-up	Results (%)		
	Type	*N*		R, MI, I	U	D
Rennie (**66**)	Sc	100	Life	11.00	89.00	—
		222	20 yrs.	38.29	61.71	—
		134	9 yrs.	52.23	47.76	—
Malamud & Render (**52**)	Sc	177	5–9½ yrs.	32.00	58.00	10.00
Rupp & Fletcher (**70**)	Sc	608	4½–10 yrs.	21.90	63.50	14.60
Hunt, Feldman, & Fiero (**37**)	DP	641§	3½–10½ yrs.	35.10	64.90	—
Cheney & Drewry (**13**)	DP	438‡	2–12 yrs.	41.00	59.00	—
Whitehead (**80**)	DP	102	5½ yrs.	51.00	43.00	6.00
Bond &	V*	578	5 yrs.	53.00	25.00	22.00
Braceland (**11**)	DP	118	5 yrs.	31.80	56.80	8.50
Gelperin (**28**)	Sc	235	5 yrs.	40.00	54.00	6.00
Guttman, Mayer-Gross, & Slater (**30**)	Sc	188	2–4 yrs.	42.80	50.90	4.10
Romano & Ebaugh (**68**)	Sc	314	Up to 4 yrs.	23.57	76.43	—
Whitehead (**80**)	DP	90	6–18 mos.	36.00	62.00	2.00
Cheney & Drewry (**13**)	DP	500†	Im	37.00	63.00	—
Rennie (**66**)	Sc	500	Im	41.08	57.34	0.58

* Includes DP, MD, GP, IM, P.som., SP, A, PN, Pa, U, PsPath., En.

† Of the 500, 486 were discharged; 5 died in the hospital and 9 remained in the hospital. These 486 were then subsequently followed up.

‡ Of this group, 10% died; 43% continued in the hospital, and 47% were living outside.

§ Of the 641, 10.8% died (69), most of them unimproved.

— Indicates either failure to report or reporting in a manner that could not be tabulated here.

ment of psychotics that studies with a brief follow-up of one year or less, such as those of Fuller, using heterogeneous groups indicate discharge rates and recovery rates of about 27%, increasing with the increase in duration of follow-up. The rates for DP cases are lower, less than 20%. On fifteen-year follow-up Fuller found a discharge rate of 35.3%, while the over-all discharge rate for all psychoses (for those not later readmitted) was 40.9%. The early studies of Bond with heterogeneous groups are similar to Fuller's. Including the young and old, the functional and the organic cases, Bond obtained a recovery rate of about 25% for five-year follow-up and a total improvement rate of about 40% (including recovery and improvement).

In Table 5, one heterogeneous study, that of Bond and Braceland, showed a recovery rate of 35% after five years and a total improvement rate (recovered and improved) of 53%. Most of the other studies, all of which have been with homogeneous groups—schizophrenics (see Table 5), have indicated that about a 40% improvement rate may be expected over a five-year period. The tremendous variation in all of these studies in respect to follow-up, duration of illness, and the like, makes strict comparisons impossible.

Outcome of Insulin, Electroshock, Metrazol and Psychosurgical Therapies (Without Controls)

The introduction of the shock therapies and then, later, of psychosurgery was heralded by interested and hopeful psychiatrists as a great advance. The initial enthusiasm for insulin, metrazol, the electroconvulsive therapies, and psychosurgery as well has, however, been giving way gradually to more caution in most circles since the mid-1940's at which time the five-year follow-up results of the shock therapies and psychosurgery began to appear.

In general, the studies on the outcome of these specific somatotherapies can be considered under two categories, depending on whether or not control groups have been used. The first group which we are discussing here simply reports the outcome of the particular treatment used. Such studies as a rule simply indicate the immediate and/or follow-up status of patients treated with the given therapy. No attempt is made to use a control group. The general implication seems to be that the patients would have been worse if the particular treatment had not been employed. A group of such studies is presented in Table 6.

TABLE 6

STUDIES ON THE OUTCOME OF SPECIFIC SOMATOTHERAPIES (WITHOUT CONTROLS)

Study	Patient		Duration of Illness	Duration of Follow-up	Type of Therapy	Results (%)		
	Type	*N*				R, MI, I	U	D
Palmer & Braceland (64)	MD-m	32	Va	Im	PNa	87.50	12.50	—
	MD-d	6	Va	Im		66.67	33.33	—
	MD-ag.d.	8	Va	Im		50.00	50.00	—
	Sc	46	Va	Im		54.60	45.40	—
	IM	3	Va	Im		66.67	33.33	—
	PN	5	Va	Im		100.00	0.00	—
Bateman & Michael (3)	Sc	416	Va	Im	I	71.90	27.60	.20
	Sc	579			Me	61.80	37.60	.30
Halpern (31)	Sc	25	Ac	Im	I	95.00	5.00	—
		31	Ch			16.13	83.87	—
		—				—	—	—
		56				48.16	51.62	
Bond & Rivers (12)	Sc	188	Va	Im	I	55.00	—	—
Epstein (17)	MD-m	13	6–28 mos.	Im	ECT	69.10	30.70	—
	MD-d	24				87.50	12.50	—
	IM	16				81.30	18.70	—
	Sc	37				43.20	56.40	—
	M	10				50.00	50.00	—
Fitzgerald (20)	DP	150	6 mos.–5 yrs.	Im	ECT	82.00	8.00	—
Impastato & Almansi (40)	MD-d	658	Va	Im	ECT	91.00	9.00	0.00
	MD-m	184				81.00	19.00	0.00
	IM	322				88.00	12.00	0.00
	PN	282				70.00	30.00	0.00
	Sc	391	Ac			76.40	23.60	0.00
	Sc	128	SAc			62.50	37.50	0.00
	Sc	412	Ch			38.00	61.70	0.00
Malzberg (54)	DP	491	Va	Im	ECT	60.10	39.18	0.00
	MD	142	Va	Im		85.80	13.40	0.70
	IM	85	Va	Im		78.80	18.80	2.40

TABLE 6 (*continued*)

Study	Patient		Duration of Illness	Duration of Follow-up	Type of Therapy	Results (%)		
	Type	*N*				R, MI, I	U	D
Hofstatter et al. (34)	Sc	66	10 yrs.*	Im	L	—	—	—
	MD	16				—	—	—
	PN	6				—	—	—
	O	12				—	—	—
	AP	100				67.00	31.00	2.00
Kindwall &	V	15	1–18 yrs.	Im	L	67.00	33.00	0.00
Cleveland (44)								
Kwalwasser &	DP	54	Va	Im	I	51.90	48.10	0.00
Robinson (45)	DP	168				81.60	17.20	1.20
Smith et al. (74)	IM	62	22 mos.	Im	ECT	90.00	8.00	1.00
	MD-d	125	8–10 mos.			87.00	13.00	0.00
	MD-m	30	Us			76.00	24.00	0.00
	Sc	27	2 wks.–2 yrs.			10.00	90.00	0.00
	Ud	20	1 mo.–10 yrs.			45.00	55.00	0.00
	PN	15	Ch			33.00	67.00	0.00
Oltman et al. (63)	Sc	91	7.4 yrs.*	Im	L	77.00	17.60	4.40
	O, AfP, Pa	16				87.50	6.20	6.20
	AP	107				78.50	15.90	4.70
Paster &	V (majority	570	Ac	Im	ECT	61.00	39.00	0.00
Holtzman (65)	Sc)	189			ECT & I	69.00	31.00	0.00
		241			I	81.00	19.00	0.00
Kane et al. (42)	Sc	106	1–50 yrs.	Im	L	42.50	—	—
	O	16						
	AP	122				64.00	—	—
Bennett (4)	IM	24	1 wk.–3 yrs.	3–18 mos.	Me	96.00	4.00	0.00
Wilson (81)	IM	19	13 mos.	Im	Me	69.00	30.00	—
				6 mos.		78.00	21.00	—
Malamud et al. (51)	IM	14	<2 yrs.	9–21 mos.	Me	79.00	21.00	—
Bond & Rivers (12)	Sc	159	Va	6 mos.	I	42.00	—	—
		141		1 yr.		42.00	—	—
		108		2 yrs.		36.00	—	—
		71		3 yrs.		31.00	—	—
		45		4 yrs.		37.00	—	—
Bennett & Wilbur (5)	IM	41	22 mos.	3–63 mos.	ECT	90.00	10.00	—
Freeman & Watts (22)	IM	108	Va	6 mos.	L	50.00	47.00	3.00
	Sc	126		9 yrs.		46.00	52.00	2.00
	Ob	51				61.00	35.00	4.00
	PN	38				63.00	37.00	0.00
	U	8				25.00	37.00	38.00
Holt (35)	DP	231	38 mos.*	1–2 yrs.	Me	61.00	39.00	0.00
MacKinnon (50)	IM	9	Us	1–3 yrs.	ECT	68.00	—	—
Morrow & King (59)	IM	45	Us	1–10 yrs.	ECT	89.00	6.00	2.00
Moure et al. (58)	Se	233	2 yrs.°	1 yr.	L			
	O	61						
	AP	294				60.80	39.10	0.00
Martin (55	Sc	239	Va	1 yr.	ECT	46.00	52.00	—
	MD	45				78.00	22.00	—
	IM	32				81.00	19.00	—
	PN	26				54.00	46.00	—
	Sc	105	Va	8 yrs.	Me	34.00	66.00	—
	MD-m	11				73.00	28.00	—
	MD-d	5				60.00	40.00	—
	MD-mix.	3				67.00	33.00	—
	IM	3				66.00	33.00	—
	PN	14				79.00	23.00	—
Stengel (75)	Sc	200	Va	1 yr.	L	33.00	57.00	10.00

— Indicates either failure to report or reporting in a manner that could not be tabulated here.

TABLE 7

STUDIES ON THE OUTCOME OF SPECIFIC SOMATOTHERAPIES (WITH CONTROLS)

Study	Patient			Duration		R_x	Results (%)			Nature of Control Group
	Type	Group	*N*	Illness	Follow-up		R, MI, I	U	D	
Taylor & Van Salzen (78)	DP	C	638	2 yrs. or less	Im	I	54.00(a)	—	—	Preshock data, own hospital
		R_x	40				62.00(a)	—	—	
Libertson (47)	Sc	C	65	9 yrs.*	Im	I	40.00	56.00	4.00	Stenographer selected at random a series of non-insulin treated patients admitted during insulin era
		R_x	165	2.4 yrs.*			62.40	36.90	.60	
Ziskind et al. (83)	AfP	C	109	9.7 mos.	Im	(58)Me	70.00	15.00	15.00	43 refused convulsive therapy 50 symptoms too mild for this R_x 16 R_x contraindicated because of physical disease
		R_x	88	6.4 mos.		(30)ECT	96.00	1.00	3.00	
Hinko & Lipschutz (32)	Sc	C	289	Us	Im		35.60	—	—	Preshock era: spontaneous remissions
		R_x	191			I	53.40	—	—	
		R_x	242			Me	37.60	—	—	
		R_x	24			ECT	45.80	—	—	
Tillotson & Sulzbach (79)	MD-d	C	68	Us	Im	ECT	13.00(d)	—	—	Comparable group without use of shock treatment
	IM	R_x	70				31.00(d)	—	—	
	RD									
Tait & Burns (77)	IM-M	C	101	Va	Im	Me and/or ECT	55.00(c)	37.00	8.00	Of the 141 control patients: 53 were admitted during the preshock era (1920–1937) 88 were admitted during the postshock era (1938–1949) but did not receive ECT for reasons unspecified
		R_x	171				92.00(c)	6.00	2.00	
	IM-Pa	C	22				57.00(c)	43.00	0.00	
		R_x	42				90.00(c)	10.00	0.00	
	O	C	18				58.00(c)	5.00	0.00	
		R_x	25				100.00(c)	0.00	0.00	
Gottlieb & Huston (29)	Sc	C	132	6–19 mos.*	1–4 yrs.	I	43.00	54.00	3.00	Control group of Malamud & Render (52)
		R_x	66				47.00	52.00	1.00	
Bond (10)	Sc	C	153	Va	Im	I	4.00	—	—	Consecutive cases from Pennsylvania Hospital followed for 5 years from date of admission. Diagnoses made by the same psychiatrist as the insulin cases. (Stayed at 16% level for 5 years.)
		R_x	125				51.00	—	—	
		C			1 yr.		16.00	—	—	
		R_x					36.00	—	—	
					2 yrs.		16.00	—	—	
							29.00	—	—	
Malzberg (53)	DP	C	1,039	Ch	1–2 yrs.	I	22.10	73.30	4.60	First admissions for DP from 7/1/35–6/30/36—postshock era
		R_x	1,039				65.40	33.40	1.30	
Ross & Malzberg (69)	DP	C	1,039	Ch	1 yr.		22.10	73.30	4.60	Same as above
		R_x	1,757			I	63.60	35.20	1.10	
		R_x	1,140			Me	36.00	63.50	.50	
Notkin et al. (61)	Sc	C	69	Va	8 mos.–2 yrs.	I	21.70	78.30	0.00	Group given intramuscular injections of physiologic salt solution at time R_x group received insulin
		R_x	100				36.00	61.00	3.00	
Lipschutz & Cavell (48)	Sc	C	100	<1 yr.	Us	I	20.00	—	—	Preshock data, own hospital. Every third Sc patient on list was used.
		R_x	35	4 yrs.*			35.00	—	—	

TABLE 7 (*continued*)

Study	Patient			Duration		R_x	Results (%)			Nature of Control Group
	Type	Group	*N*	Illness	Follow-up		R, MI, I	U	D	
Notkin et al. (**62**)	Sc	C	71	>18 mos.*	6–18 mos.	Me	8.40	91.60	0.00	Group given a sterile physiologic solution of sodium chloride and kept on same wards with R_x group
		R_x	100				18.00	81.00	1.00	
Libertson (**47**)	Sc	C	65	9 yrs.*	7–9 mos.	I	40.00	56.00	4.00	Stenographer selected at random a series of non-insulin treated patients admitted during the insulin era
		R_x	165	2.4 yrs.*			40.00	57.50	2.40	
Ziskind et al. (**82**)	AfP	C	47	8 mos.*	10 mos.*	Me	72.00	17.00	11.00	15 refused metrazol
		R_x	38	7 mos.*			92.00	5.00	3.00	22 had symptoms too mild for this R_x 10 R_x contraindicated because of physical disease
Tillotson & Sulzbach (**79**)	MD-d	C	68	Us	18–45 mos.	ECT	13.00	—	—	Comparable group without the use of shock treatment
	IM	R_x	70				57.00	—	—	
	RD									
Ziskind et al. (**83**)	AfP	C	109	9.7 mos.	40 mos.	(58)Me	70.00	15.00	15.00	43 refused metrazol
		R_x	88	6.4 mos.		(30)ECT	89.00	7.00	4.00	50 symptoms too mild for R_x 16 R_x contraindicated because of physical disease
Holt (**35**)	DP	C	61	<1 yr.	1 yr.	Me	77.00	23.00	—	Not stated
		R_x	62				90.00	10.00	—	
		C	11	1–3 yrs.			55.00	45.00	—	
		R_x	32				66.00	34.00	—	
		C	32	>3 yrs.			47.00	53.00	—	
		R_x	29				52.00	48.00	—	
Hinko & Lipschutz (**32**)	Sc	C	289	Us	5 yrs.		23.80(a)	—	—	Preshock era: spontaneous remissions
		R_x	191			I	29.30(a)	—	—	
		R_x	242			Me	26.00(a)	—	—	
		R_x	24			ECT	33.00(a)	—	—	
Huston & Locher (**38**)	MD-d	C	80	6.3 mos.*	82 mos.	ECT	79.00	9.00	12.00	Preshock records, own hospital
		R_x	74	6.8 mos.*	36 mos.		88.00	10.00	2.00	
Huston & Locher (**39**)	IM	C	93	54 mos.*	Up to 18 mos.	ECT	46.00	18.00	36.00	Preshock records, own hospital
		R_x	61		6–48 mos.		100.00	—	—	
Fishbein (**19**)	IM	C	61	Us	Us	ECT	91.70	8.30	—	Preshock records, own hospital
		R_x	347				88.10	11.80	—	
Friedman et al. (**23**)	Sc	C	100	Ch	2 yrs.	L	2.00	—	2.00	Ori-inally selected for lobotomy, but family permission not granted. *N.B.* 21% of this control group received shock therapy during the follow-up period.
		R_x	254				57.80	38.20	3.90	
Finiefs (**18**)	Sc	C	446	<1 yr. to	5 yrs.		34.50	—	—	Received no special R_x beyond ordinary nursing care, routine hospital care
		R_x	563	3 yrs.		ECT	39.00	—	—	
						I	54.20	—	—	
						I*	54.30	—	—	

Note.—C indicates control group; R_x indicates group receiving treatment. —Indicates either failure to report or reporting in a manner that could not be tabulated here.

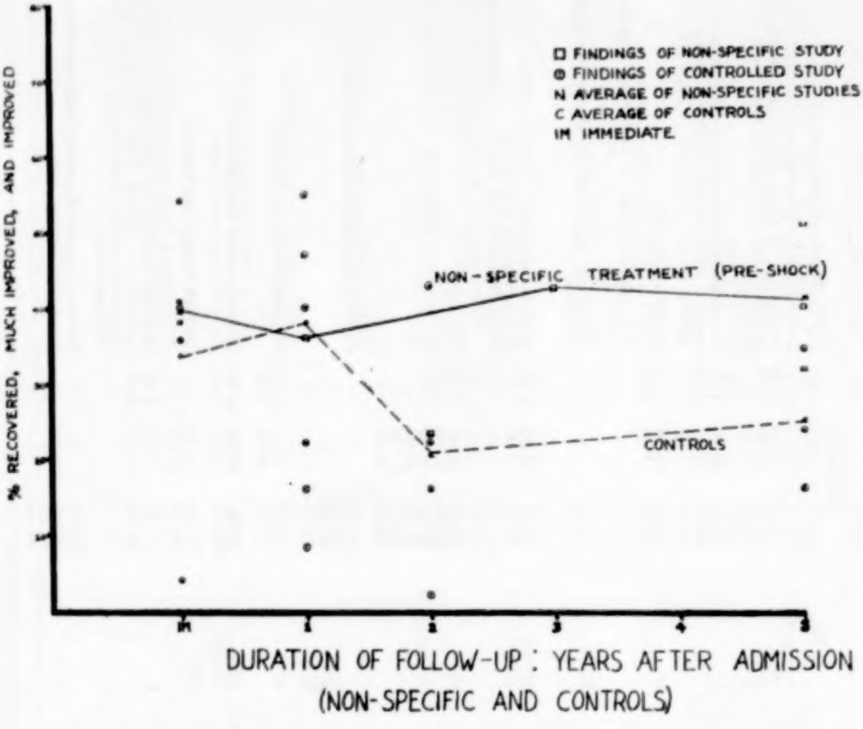

FIG. 2. OUTCOME OF MENTAL PATIENTS GIVEN NONSPECIFIC TREATMENT (PRESHOCK) AND OF THOSE USED AS CONTROLS (SCHIZOPHRENICS)

Results of Comparative Studies with Treated and Control Groups

While the studies just described were without controls, others have made an attempt to compare untreated groups with those treated with a specific somatotherapy. In these investigations control groups have been assembled from current cases for the particular research in progress, or control data have been obtained from the preshock records of the investigator's own practice or hospital. Table 7 presents a representative group of studies containing treated groups as well as untreated control groups.

Graphic Analysis of Previously Tabulated Results on Somatotherapies

In order to analyze the results of the previously tabulated studies, specifically for schizophrenia, a series of graphs was made. Figure 2 shows the percentage of schizophrenics recovered, much improved, and improved at follow-up intervals up to 5 years after admission. The findings of individual studies and control studies have been shown in the graph, but the lines are drawn through the averages for nonspecific treatment at each follow-up interval in the one case, and through the averages for the con-

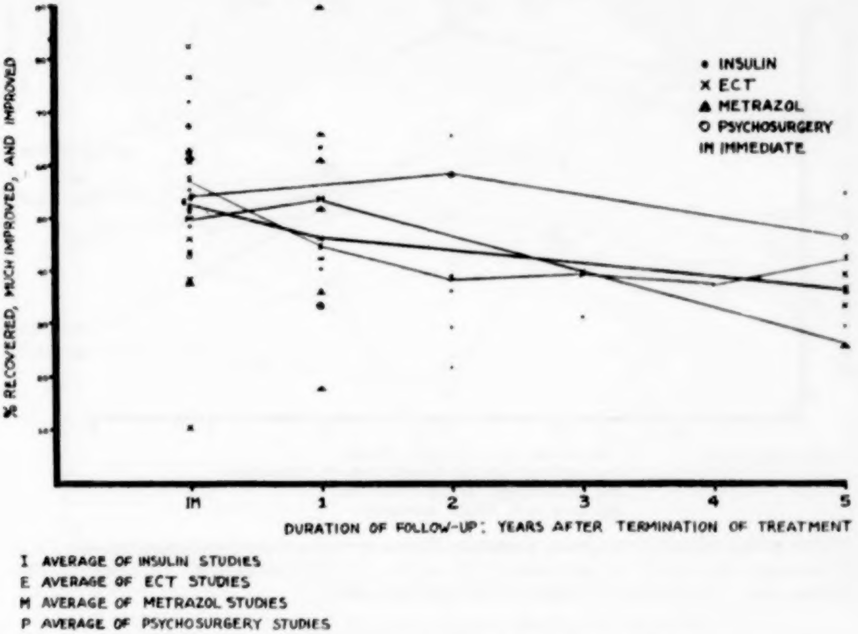

FIG. 3. OUTCOME OF MENTAL PATIENTS TREATED WITH VARIOUS SOMATOTHERAPIES (SCHIZOPHRENICS)

trols at each follow-up interval in the other. One exception should be pointed out. It will be noted that the nonspecific treatment study at the two-year level (a single study—23.57% recovered ànd improved) drops unusually low and out of line. The point is irldicated in the graph, but omitted from the line of averages.

In general the trend is for the nonspecific treatment groups to show a recovered, much improved, and improved rate of about 40% at the end of five years, a finding in conformity with those of Bond and Fuller. It is interesting to note, however, that with the single exception of the one-year follow-up level, the controls are always poorer than the nonspecific groups, showing a recovered, much improved, and improved rate of only about 25% at the end of five years. One may ask why the controls used in the shock era should be so different. One reason may be that in the shock era only poorer patients, that is, those with unfavorable prognoses, were available as controls, other patients being given the benefit of the specific treatments. There is also a second possibility, namely, that chronic, deteriorated cases may have been selected as controls. In any case the outcome for the controls is strikingly different from that for those given nonspecific treatment in the 1930's. The general findings of these studies of nonspecific treatments and spontaneous remissions should be kept in mind as we turn now to the results that have been reported for the specific somatotherapies—insulin,

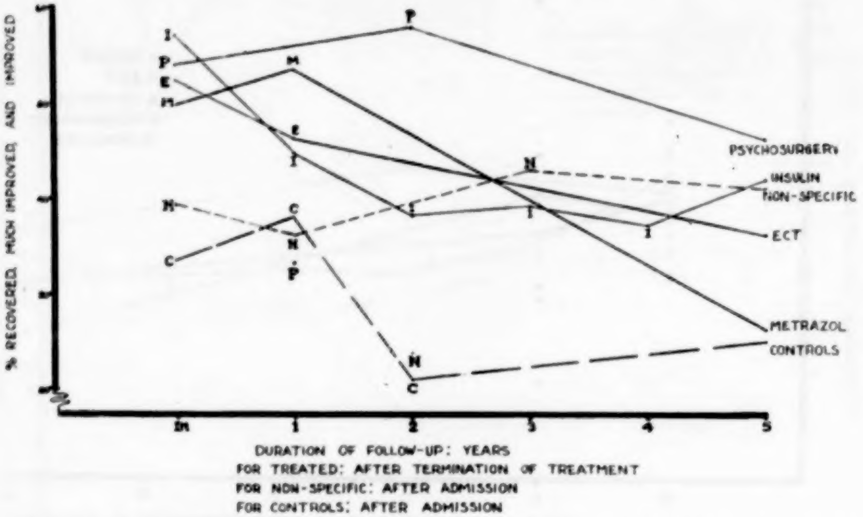

Fig. 4. Outcome of Mental Patients Treated with Various Somatotherapies as Compared with Outcome of Mental Patients Given Nonspecific Treatment (Preshock) and Those Used as Controls (Schizophrenics)

metrazol, electroconvulsive therapy, and psychosurgery.

Next, the outcome of schizophrenics treated by the various somatotherapies, that is, insulin, metrazol, electroshock, and psychosurgery, was analyzed. These results are presented in Fig. 3. The percentage of schizophrenics recovered, much improved, and improved in the individual studies at varying follow-up intervals after the termination of treatment, up to five years, is shown in this graph. As in the previous graph, the averages at each interval of follow-up have been computed. The insulin average is indicated as *I*, electroshock as *E*, metrazol as *M*, and psychosurgery as *P*. The averages for each treatment at each interval constitute the points through which the lines are drawn to represent the outcome for each type of treatment. The graph reveals considerable variability and overlap among the different types of somatotherapies as well as a dearth of studies for longer follow-up periods. Most studies on the outcome of therapy report only immediate outcome and very few go beyond the one-year period. At five years all outcome results for the somatotherapies are poorer than immediate outcome.

In Fig. 4 the lines shown in Fig. 2 and 3 have been combined. Figure 4 shows the average outcome results for schizophrenics treated by each somatotherapy, for the nonspecific treatment and for the controls. The average results for all insulin studies are plotted as *I* at each interval, metrazol as *M*, electroshock as *E*, and psychosurgery as *P*, nonspecific as *N*, and the controls as *C*. The lines drawn through these average points at the various follow-up intervals indicate in all cases better immediate outcome for treated patients (about 50–60%) than for the nonspecific and

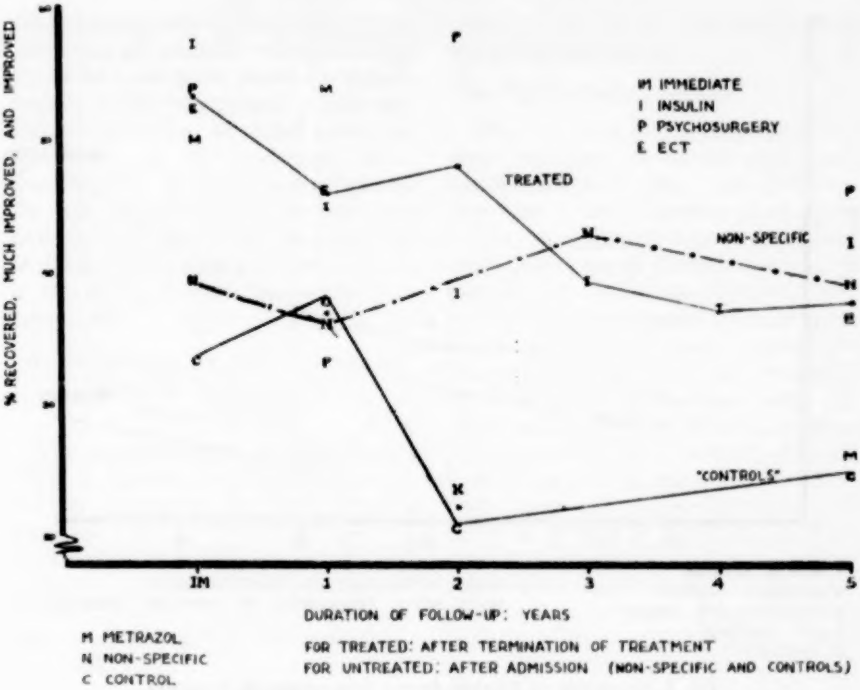

FIG. 5. TREATED VS. UNTREATED PATIENTS (SCHIZOPHRENICS)

controls, but this advantage is not maintained for the treated group on follow-up after five years. Whereas the nonspecific group never shows such striking recovery and improvement rates, the treated groups show more relapses with time, dropping toward the nonspecific rate of about 40% after 5 years following treatment, but never dropping as low as the controls.

When all the treatments are averaged into a single line and when the nonspecific and controls are included, we get the results shown in Fig. 5. For comparison purposes we have also drawn in the nonspecific and control averages. In the graph the letters *I*, *P*, *E*, *M*, *N*, and *C* indicate the average results for all studies of insulin, psychosurgery, electroshock, metrazol, nonspecific treatment, and controls at a given follow-up period (for the somatotherapies from the date of termination of treatment, and for the nonspecific and controls from the day of admission). It will be immediately observed that the recovered, much improved, and improved rate among treated schizophrenics tends to decline as the period after treatment increases. For the untreated the course is more variable, but when the nonspecific treatments are considered alone, they are slightly better than the treated after 5 years, with a much more even course than the treated during the five-year period.

Although from these results it would appear that somatotherapy

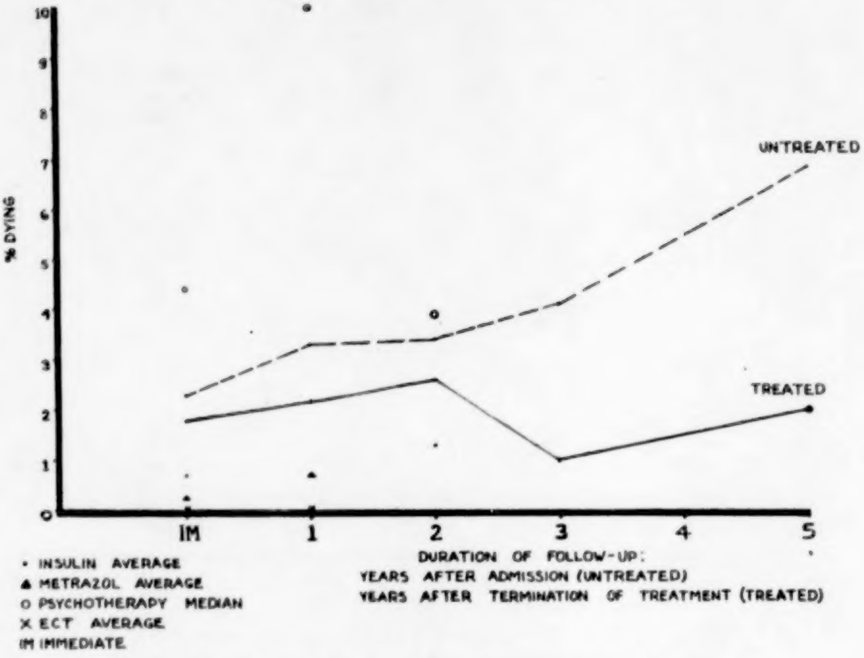

FIG. 6. INCIDENCE OF DEATHS AMONG SCHIZOPHRENIC PATIENTS

for schizophrenics offers little advantage when long-term follow-up results are considered, yet another aspect of the problem needs to be considered, that is, the incidence of deaths among the treated and the untreated. Figure 6 presents the percentage of schizophrenics dying in the untreated and the treated groups during follow-up periods, in the case of the treated from the date of the termination of treatment, and in the case of the untreated (nonspecific and control patients) from the day of admission. The average deaths on follow-up for patients who had been treated with insulin, metrazol, psychosurgery, and electroshock are shown in Fig. 6. The lines representing average nonspecific and control deaths are also shown. In this graph we can see immediately that the death rate rises steadily in the untreated groups (nonspecific and controls). Fewer patients die in the treated than in the untreated groups. (It should be noted too that psychosurgery contributes considerably more to the death rate than the other therapies.) The saving of life which treatment offers cannot be overlooked, even in the face of failure to produce high recovery and improvement rates.[3]

The data presented in these graphs do not permit us to draw any definite conclusions as to the relative merits of any one specific therapy in the treatment of schizophrenia, since the

[3] The conclusion that the death rate among the treated is lower than that among the controls is only tentative, since it was impossible to obtain information regarding age-specific and sex-specific death rates for the treated and the control groups.

evidence is scattered and since, moreover, not all studies on somatotherapy which we have found in the literature could be graphed in this way (largely because of insufficient information as to follow-up, etc.); nevertheless, they shed some light on the inadequacies of the studies that have been done and they indicate the need for improvement of the research in the evaluation of the somatotherapies (**33**).

An Appraisal of the Foregoing Studies in Terms of Methodology

The methodology of the studies listed in Tables 6 and 7 will be considered in terms of the four essentials previously set by one of the present authors (**84**) as minimum essentials of adequate research design: homogeneity, control groups, follow-up, and specific criteria for evaluating outcome.

Homogeneity

It can be seen from both Tables 6 and 7 that some investigators continue to include all types of mental diseases in their investigations without specification. While this practice is acceptable if separate outcome results are presented, some give only the total results (**44**, **58**). Very frequently acute and chronic cases are included in the same study, and total, rather than separate, outcome results are reported. In most cases the length of the illness before treatment is not even mentioned. While most investigators usually give adequate identifying data as to the age, sex, and number of patients, an occasional study does not even clearly specify the disease entity under investigation. From the point of view of research design in general, however, homogeneity is probably the feature least open to criticism in modern evaluations of the somatotherapies.

The Use of Control Groups

This criterion of research design is often not met at all or only very poorly satisfied (**15**). In Table 6 there are a large number of studies, no one of which has used any controls. This type of study is commonly found in the literature. The immediate or follow-up results of treatment are presented, the basic assumption being that the patients studied would have been worse if untreated. In many of the studies which have employed controls (see Table 7), the difficulty lies in the nature of the controls used. In some investigations old studies such as Bond's have been cited and used as norms against which to compare shock results. These data ought not to be used as standards or base lines since, as we have seen, they were derived from total hospital populations. Moreover, diagnostic criteria have changed and probably the character of the patient population has changed, too, as a consequence of mental health education, interest in psychiatry, and increased use of the specific therapies in private practice as well as in hospitals. The use of the control data of other workers and other hospitals as reported in the literature, no matter how recent, is valueless because of the discrepancy in diagnostic criteria. Yet this practice has not been abandoned by investigators (**29**).

Several investigators, appreciating the inadvisability of employing control data from other studies, have assembled controls for their particular research. But a review of Table 7 will reveal that this practice is not without its difficulties. Often such control groups include patients in whom the treatment is contraindicated (**82**, **83**), whose symptoms are too mild for treatment by the particular therapy under investigation (**82**, **83**), or who are chronic and deteriorated cases. Some control groups include an assortment of all these different types (**82**, **83**). Occasionally, the nature of the control group is merely described as a comparable group but without shock treatment. No details are given (**79**). At times, control cases are selected at random by a secretary (**47**). Very often the control groups are not given the same amount of motivation as the treated groups, so that their morale may be lower during the observation period. On that account the groups may not be strictly comparable. Sometimes the control groups or some of their members are given other treatment during the period of follow-up (**23**).

Needless to say these practices render the findings of such research worthless. The problem of establishing controls in psychiatric research has provoked much discussion in the literature, some writers taking points of view that are diametrically opposed. Curran (**14**) has indicated that obtaining controls in psychiatry is an arduous task, for it is not easy to assemble groups of patients who can be validly compared. In discussing electric shock treatment and its results, Reynolds (**67**) has taken a similar view. Moreover, Curran argues that it is never possible to get untreated groups, for he feels that it is not possible to determine the nature of a reaction without altering that reaction at least to some extent, since in any medical examination some impressions as well as some recommendations are made.

The selection of controls is no easy matter. First of all, our ignorance of the causes of mental diseases militates against matching controls and treated cases with certainty. Second, it is not right, ethically speaking, to withhold certain treatments, just as it would be wrong to administer certain untried treatments, for research purposes only. One feature which has helped the research worker is the failure of some families to consent to a specific therapy for which a particular patient has been selected. Unfortunately, the attitude of the family may be a factor in the final release of the patient, thus producing a discrepancy between the treated and untreated groups (**22**).

Occasionally the objection is raised that the morale of the control groups is lowered by failure to receive the specific therapies. This objection has been met by Notkin (**61**, **62**), for example, who gave intramuscular injections to control group patients while the treated patients received insulin. This objection may also be adequately met by providing the control group with "total push." In this connection, it is interesting to note that Mettler (**57**) has urged that three comparable groups be used, if possible: (*a*) one group should be followed to investigate the degree of spontaneous improvement which may occur; (*b*) another should receive the specific therapy to be evaluated; (*c*) if a third group exists, it should be subjected to all the nonspecific aspects of the therapy to be evaluated. Theoretically Mettler's suggestion is sound, but in practice it further complicates the problem by requiring the selection of three comparable groups instead of two. Since the use of controls is imperative in any sound experimentation, this important criterion continues to be one of the troublesome features of measuring the effectiveness of the somatotherapies.

One method is to lay down a minimum number of variables on which the treated and the untreated groups should be comparable. Among such fundamental variables are: age, sex, age at onset of illness, type of onset, type of illness, type of treatment, duration of follow-up. Once comparability is attained on these fundamental variables, additional variables on which the control and treated groups may differ can be controlled by analysis of covariance or similar methods. A list of variables which may be important in determining outcome is currently under study in an investigation of prognosis (Zubin, J., Peretz, D. and Ossipow, S., Psychiatric Prognosis—in preparation).

Follow-up

In both the uncontrolled and the controlled studies, the *duration* of follow-up is frequently unspecified. In some studies it varies from patient to patient. The duration of follow-up is a prime consideration, for immediate evaluations of the outcome of therapy or evaluations based upon short follow-up periods make no allowance for the possibility of later relapses. Thus the recovery rates in studies that give the patients' status immediately or shortly after termination of treatment are spuriously high. Menninger (**56**) and Alexander (**2**) have for this reason both emphasized the importance of the time factor in evaluation studies by stressing how different therapeutic results may appear after varying intervals following treatment. In general, long-term follow-up, preferably a period of five years, is the ideal. Only a few studies meet this requirement (**18**, **34**, **39**). It is of course difficult and expensive to keep each member of a population under observation for so long a period. Furthermore, populations are not static. There are remissions, deaths, and patients lost to the study because of removal from the community. Dorn (**16**) has cautioned that the validity of a follow-up study is very questionable if each individual is not followed for the maximum possible duration. Naturally it is understandable that all patients cannot always be followed. The investigator should mention how many could be traced after a given follow-up period, but this information is seldom supplied. More commonly, conclusions are based on the number for whom information is available with no reference to missing cases. Obviously this gives rise to biased results.

While we have not indicated in our tables what *method* of follow-up was employed in each study, it should be mentioned here that in a large number of studies the procedure is rather haphazard. Rarely is it uniform for all patients. As a rule, questionnaires are sent to patients or to their families. Social service or psychiatric interviews are given to some, but not to all. Occasionally in the same study some patients are followed up in every possible way, some in only one way, failing which, the patient is lost to the study. The variation in follow-up methods for subjects in the same investigation is considerable.

In general the interview is considered preferable to the questionnaire by all workers. Ideally, for re-examination, the psychiatrist as well as the social worker should see the patient, preferably the same psychiatrist and social worker responsible for original examination. Patients' reports of their own status or those obtained from their families, important as they are, should not constitute the sole basis for evaluation.

Another problem of follow-up which is critical is the treatment of deaths.

Some workers such as Karagulla (**43**) exclude deaths; others such as Slater (**72**) include them in computing the improvement rate. In several studies deaths are not even reported. It has been suggested (**85**) that Jerzey Neyman's (**21**) mathematical models can provide the real answer to this problem through the computation of net improvement rates from which the influence of deaths and relapses has been eliminated.

Criteria of Evaluation

In reviewing the studies on the outcome of therapy one cannot help being impressed with the number of outcome categories which research workers have been able to devise. Recovery may be variously expressed as complete recovery or social recovery, while improvement may be described in terms of improved, much improved, slightly improved, little improvement, and the like. In our tabulations we have grouped these various headings under the more general captions, Recovered and Improved, Unimproved and Dead. Not only do different investigators use different categories, but they define the same categories differently. Objective terms are needed to express changes in the patient's condition after treatment. Objectivity, however, is difficult to achieve wherever the nature and extent of improvement are obscure, as is usually the case in psychiatric disorders. One must determine whether a specific change in a patient's condition signified improvement. The extent of error in judging the presence of improvement and its degree should be indicated. Gjessing (**71**) in 1938, after noting the impossibility of keeping meaningful statistics until it is known what a given worker means by the terms "recovered," "improved," "much improved," urged that international standards be defined for these terms. One can only regret, after a review of the current literature, that Gjessing's suggestion was never implemented.

Because of the lack of uniformity and objectivity in reporting outcome, exact comparison of the various studies is difficult. It appears that the only uniform category for all studies is the Dead, although in some studies the number of the dead is not listed separately, being included under other categories according to the status at the time of death. Where deaths are reported, the figures are often presented apologetically, at times with reassuring observations to the effect that death was not really due to the treatment.

In reporting outcome, workers have demonstrated considerable variation in the use of numbers or percentages of patients. Some authors use parole or discharge as the criterion of evaluation. In spite of its limitations, as previously mentioned, this criterion does seem to be about the most satisfactory, since it reduces the classification to a dichotomy. The patient is either *in* or *out*. Where there are multiple classifications, they are invariably based upon subjective clinical evaluations with consequent increase in the possibility of error. More recently, as in the Brain Research Project at the New York State Psychiatric Institute, objective rating scales have been employed with a scale of numbers ranging from 1 to 5 (excellent . . . most unfavorable outcome). This lends itself more readily to statistical evaluation without in fact eliminating the subjective character of the rating. Such a rating scale, universally used, would make for uniformity of outcome categories, and by reducing the present variation from study to study, would facilitate comparisons of various investigations.

Need for Planned Research

Various suggestions for planned research (**73**) have been formulated from time to time by different investigators. Thus Luff (**49**) urged close cooperation between workers in different mental hospitals, pointing to the kind of cooperative inquiry employed successfully in cancer research as an example of what could be done. In 1937 Luff suggested that mental hospitals keep standard records and institute follow-up systems such as are maintained for cancer study. Systematic research seems to be just as necessary in the field of mental disease as for physical disease. After reviewing the literature for the present paper, however, the writers are left with the impression that a large percentage of the articles on the evaluation of the somatotherapies was inspired by the mere fact that certain data had been collected. An article in a psychiatric journal seemed a natural way to make use of them. In other words, the planning often seems to have come last. This is probably the reason why a large number of investigations reported in the literature on evaluation of therapy appear purposeless, disorganized, and poorly executed, despite their impressive arrays of statistics. Various complaints have been raised against these surveys of therapeutic results. Some critics, like Israel and Johnson (**41**), have felt that prevailing statistics do not accurately portray the really hopeful prognosis for the mentally ill. Others have complained that present statistics are too optimistic.

As early as 1930, Sakel (**71**), alarmed by the already numerous statistical reports on shock therapies, cautioned that psychiatrists should not place too much reliance on statistics because of the lack of knowledge as to the true nature of the mental diseases which they were studying, especially schizophrenia. In his opinion most of the researchers were dealing with symptoms, for "they had no test to define the nature of schizophrenia, and therefore they could not set up a test of 'cured' and 'not cured'." In like vein, Alexander (**1**) has complained about "the delirium of numbers" in personality research. Similarly, Lewis (**46**) in discussing the status of shock therapy in 1943 emphasized the disagreement in results and reminded his colleagues that statistical manipulation of unreliable data is fruitless. Moreover, there seems to be a clear division between one group of workers who rely on clinical judgment and another group who enlist statistics in appraising the results of the psychiatric therapies. In any case, research evaluating the outcome of therapy needs careful planning.

Summary

This review of the literature evaluating the somatotherapies reveals that a large number of studies have been inadequately planned and poorly designed. Several serious defects have been observed. Among them are: (*a*) *lack of homogeneity* of patients studied in respect to diagnostic classification, age, duration of illness, and follow-up; (*b*) too brief, poorly executed, or inadequately reported *follow-up;* (*c*) *lack of controls* or poorly selected controls; (*d*) inadequate, ill-defined, or unspecified *criteria for evaluating outcome;* (*e*) failure to report *deaths*, especially for follow-up studies, or inclusion of the dead under the category representing their status at the time of death.

In spite of their individual limitations these studies, taken in the aggregate, have demonstrated short-term advantages but have not demonstrated definitely significant advantages for the specific somatotherapies in the long run. Our review of the literature, however, has revealed the following facts:

1. Where only immediate outcome is reported, there seems to be a distinct advantage for treated groups as compared with untreated ones. Their stay in the hospital is reduced. The death rate is apparently lower for the treated. Such results should not be underestimated, for they may mean that suicides and deaths from inanition have been reduced among depressed patients; that human suffering has been alleviated and that private and state funds for hospital care have been saved even though relapses do occur. The somatotherapies help to save human life.

2. Long-term follow-up studies have not generally shown better results for the treated over the untreated in terms of recovery and improvement. The recovery rate still hovers around 35% to 40%. More patients tend to recover in nonspecific groups (and even in the control groups) after five years, whereas more relapses occur among the treated patients.

3. Generally speaking, the specific somatotherapies have seemed to work better with patients whose illness is of short duration (we have not taken up the problem of duration of illness in this paper). Therefore, it may be that these therapies merely accelerate the improvement process in those who would have recovered spontaneously anyway. Even this, however, must be recognized as a real advantage.

4. It is likely that if better prognostic indicators for the efficacy of each therapy were available, the results for each of the therapies would excel the outcome of the nonspecific therapies.

Finally, the present analysis of data in the literature has indicated that studies evaluating therapy must contain the following minimal information, if adequate comparisons are to be made: sex, age at onset of illness, type of onset (sudden or insidious), age at the time of treatment, type of treatment, duration of follow-up, and an accurate report of deaths. In comparative studies the selection of controls matched on the above-mentioned factors is a necessity. Future research on the evaluation of the outcome of therapy should be designed to meet these requirements and should become more exact (**36**).

REFERENCES

1. Alexander, F. Evaluation of statistical and analytical methods in psychiatry and psychology. *Amer. J. Orthopsychiat.*, 1934, **4**, 433–448.
2. Alexander, G. H. "Shock" therapies: a method of more accurate estimation of their therapeutic efficacy. *J. nerv. ment. Dis.*, 1944, **99**, 922–924.
3. Bateman, J. F., & Michael, N. Pharmacological shock treatment of schizophrenia: a statistical study of results in the Ohio State Hospitals. *Amer. J. Psychiat.*, 1940, **97**, 59–67.
4. Bennett, A. E. Metrazol therapy in affective psychoses. *Amer. J. med. Sci.*, 1939, **198**, 695–701.
5. Bennett, A. E., & Wilbur, C. B. Convulsive shock therapy in involutional states after complete failure with previous estrogenic treatment. *Amer. J. med. Sci.*, 1944, **208**, 170–176.
6. Bond, E. D. A review of the five-year period following admission in 111 mental patients. *Amer. J. Insanity*, 1921, **77**, 385–394.
7. Bond, E. D. Results in 251 cases five

years after admission to a hospital for mental diseases. *Arch. Neurol. Psychiat.*, 1921, **6**, 429–439.

8. Bond, E. D. Follow-up work in mental and surgical cases. *Amer. J. Psychiat.*, 1923, **2**, 445–450.
9. Bond, E. D. Under-estimation of good results in mental diseases. *J. Amer. med. Ass.*, 1925, **85**, 503–505.
10. Bond, E. D. Continued follow-up results in insulin shock therapy and in control cases. *Amer. J. Psychiat.*, 1941, **97**, 1024–1028.
11. Bond, E. D., & Braceland, F. J. Prognosis in mental disease. *Amer. J. Psychiat.*, 1937, **94**, 263–274.
12. Bond, E. D., & Rivers, T. D. Further follow-up results in insulin-shock therapy. *Amer. J. Psychiat.*, 1942, **99**, 201–202.
13. Cheney, C. O., & Drewry, P. H. Results of non-specific treatment of dementia praecox. *Amer. J. Psychiat.*, 1938, **95**, 203–218.
14. Curran,. D. The problem of assessing psychiatric treatment. *Lancet*, 1937, **2, 1005–1009.**
15. David, H. P. A critique of psychiatric and psychological research on insulin treatment in schizophrenia. *Amer. J. Psychiat.*, 1954, **110**, 774–776.
16. Dorn, H. F. Methods of analysis for follow-up studies. *Hum. Biol.*, 1950, **22**, 238–248.
17. Epstein, J. Electric shock therapy in the psychoses. *J. nerv. ment. Dis.*, 1943, **98**, 115–129.
18. Finiefs, L. A. The results of treatment of a thousand cases of schizophrenia. *J. ment. Sci.*, 1948, **94**, 575–580.
19. Fishbein, I. L. Involutional melancholia and convulsive therapy. *Amer. J. Psychiat.*, 1949, **106**, 128–135.
20. Fitzgerald, O. Experiences in the treatment of depressive states by electrically induced convulsion. *J. ment. Sci.*, 1943, **89**, 73–80.
21. Fix, E., & Neyman, J. A simple stochastic model of recovery, relapse, death and loss of patients. *Hum. Biol.*, 1951, **23**, 205–241.
22. Freeman, W., & Watts, J. W. Prefrontal lobotomy, a survey of 331 cases. *Amer. J. med. Sci.*, 1946, **211**, 1–8.
23. Friedman, S., Moore, B. E., Ranger, C. O., & Russman, C. A progress study of lobotomized and control patients. *Amer. J. Psychiat.*, 1951, **108**, 10–18.
24. Fuller, R. G. Expectation of hospital life and outcome for mental patients on first admission. *Psychiat. Quart.*, 1930, **4**, 295–323.
25. Fuller, R. G. Readmissions in the hospital history of mental patients during eighteen years following first admissions. *Psychiat. Quart.*, 1931, **5**, 53–67.
26. Fuller, R. G., & Johnston, M. The duration of hospital life for mental patients. *Psychiat. Quart.*, 1931, **5**, 552–582.
27. Fuller, R. G. What happens to mental patients after discharge from hospital. *Psychiat. Quart.*, 1935, **9**, 95–104.
28. Gelperin, J. Spontaneous remissions in schizophrenia. *J. Amer. med. Assoc.*, 1931, **112**, 2393–2395.
29. Gottlieb, J. S., & Huston, P. E. Treatment of schizophrenia: follow-up results in cases of insulin shock therapy and in control cases. *Arch. Neurol. Psychiat.*, 1943, **49**, 266–271.
30. Guttman, E., Mayer-Gross, W., & Slater, E. T. O. Short-distance prognosis of schizophrenia. *J. Neurol. Psychiat.*, 1939, **2**, 25–34.
31. Halpern, F. G. Insulin shock treatment of schizophrenia. *Amer. J. Psychiat.*, 1940, **96**, 1153–1165.
32. Hinko, E. N., & Lipschutz, L. S. Five years after shock therapy. A preliminary report. *Amer. J. Psychiat.*, 1947, **104**, 387–390.
33. Hoch, P. H. Progress in psychiatric therapies. *Amer. J. Psychiat.*, 1955, **112**, 241–247.
34. Hofstatter, L., Busch, A. K., Clancy, J. F., & Smolik, E. A. The results of surgical treatment in 100 cases of chronic mental illness. *Sth. med. J., Bgham*, 1945, **38**, 604–607.
35. Holt, W. L. Practical value of convulsive shock therapy research. *Dis. nerv. Syst.*, 1947, **8**, 112–117.
36. Hopkins, C. E. Psychiatry, science and statistics—a review of current trends. *Psychiat. Quart.*, 1956, **30**, 1–14.
37. Hunt, R. C., Feldman, H., & Fiero, R. P. "Spontaneous" remissions in dementia praecox. *Psychiat. Quart.*, 1938, **12**, 414–425.
38. Huston, P. E., & Locher, L. M. Manic depressive psychosis. Course when treated and untreated with electric shock. *Arch. Neurol. Psychiat.*, 1948, **60**, 37–48.
39. Huston, P. E., & Locher, L. M. Involutional psychosis. Course when untreated and when treated with electric shock. *Arch. Neurol. Psychiat.*, 1948, **59**, 385–394.

40. Impastato, D. J., & Almansi, R. J. A study of over 2000 cases of electrofit-treated patients. *N. Y. St. J. med.*, 1943, **43**, 2057–2064.

41. Israel, R. H., & Johnson, N. A. New facts on prognosis in mental disease. *Amer. J. Psychiat.*, 1948, **104**, 540–545.

42. Kane, W. J., Hurdum, H. M., & Schaerer, J. P. Prefrontal lobotomy. *Arch. Neurol. Psychiat.*, 1952, **68**, 205–212.

43. Karagulla, S. Evaluation of electric convulsion therapy as compared with conservative methods of treatment in depressive states. *J. ment. Sci.*, 1950, **96**, 1060–1092.

44. Kindwall, J. A., & Cleveland, D. Prefrontal lobotomy: fifteen patients before and after operation. *Amer. J. Psychiat.*, 1945, **101**, 749–755.

45. Kwalwasser, S., & Robinson, L. C. General survey of insulin-treated patients after 5 years. *Psychiat. Quart.*, 1949, **23**, 672–690.

46. Lewis, N. D. C. The present status of shock therapy of mental disorders. *Bull. N. Y. Acad. Med.*, 1943, **19**, 227–244.

47. Libertson, W. A. A critical analysis of insulin therapy at Rochester State Hospital. *Psychiat. Quart.*, 1941, **15**, 635–647.

48. Lipschutz, L. S., & Cavell, R. W. Evaluation of pharmalogic shock treatment. *Arch. Neurol. Psychiat.*, 1939, **42**, 365–366.

49. Luff, M. C. Assessment of psychiatric treatment. *Lancet*, 1937, **2**, 1103.

50. MacKinnon, A. L. Electric shock therapy in a private psychiatric hospital. *Canad. med. Ass. J.*, 1948, **58**, 479.

51. Malamud, W., Sands, S. L., & Malamud, I. The involutional psychoses: a sociopsychiatric follow-up study. *Psychosom. Med.*, 1941, **3**, 410–426.

52. Malamud, W., & Render, N. Course and prognosis in schizophrenia. *Amer. J. Psychiat.*, 1939, **95**, 1039–1057.

53. Malzberg, B. Outcome of insulin treatment of one thousand patients with dementia praecox. *Psychiat. Quart.*, 1938, **12**, 528–553.

54. Malzberg, B. The outcome of electric shock therapy in the New York Civil State Hospitals. *Psychiat. Quart.*, 1943, **17**, 154–163.

55. Martin, P. A. Convulsive therapies: review of 511 cases at Pontiac State Hospital. *J. nerv. ment. Dis.*, 1949, **109**, 142–157.

56. Menninger, W. C. An evaluation of metrazol treatment. *Bull. Menninger Clin.*, 1940, **4**, 95–104.

57. Mettler, F. A. A comparison between various forms of psycho-surgery. *N. Y. St. J. Med.*, 1949, **49**, 2283–2286.

58. Moore, B. E., Simon, B., Friedman, S., & Ranger, C. O. Psychosurgery: successes and failures following frontal lobotomy. *N. Y. St. J. Med.*, 1949, **49**, 2263–2273.

59. Morrow, J. K., & King, J. P. Follow-up studies of shock-treated patients. *Amer. J. Psychiat.*, 1949, **105**, 815–820.

60. Neumann, E., & Finkenbrink, F. Statische Untersuchungen über die Spontanremissionen bei Schizophrenien. *Allg. Z. Psychiat.*, 1939, **111**, 17–46.

61. Notkin, J., Niles, C. E., DeNatale, F. J., & Wittman, G. A comparative study of hypoglycaemic shock treatment and control observation in schizophrenia. *Amer. J. Psychiat.*, 1939, **96**, 681–688.

62. Notkin, J., DeNatale, F. J., Niles, C. E., & Wittman, G. Comparative study of metrazol treatment and control observations of schizophrenia. *Arch. Neurol. Psychiat.*, 1940, **44**, 568–577.

63. Oltman, J. E., Brody, B., Friedman, S., Green, W. F. Frontal lobotomy: clinical experience with 107 cases in a state hospital. *Amer. J. Psychiat.*, 1949, **105**, 742–751.

64. Palmer, H. D., & Braceland, F. J. Six years experience with narcosis therapy in psychiatry. *Amer. J. Psychiat.*, 1937, **94**, 37–57.

65. Paster, S., & Holtzman, S. C. A study of one thousand psychotic veterans treated with insulin and electric shock. *Amer. J. Psychiat.*, 1949, **105**, 811–814.

66. Rennie, T. A. C. Follow-up study of five hundred patients with schizophrenia admitted to the hospital from 1913 to 1923. *Arch. Neurol. Psychiat.*, 1939, **42**, 877–891.

67. Reynolds, W. W. Electric shock treatment. Observations on 350 cases. *Psychiat. Quart.*, 1945, **19**, 322–333.

68. Romano, J., & Ebaugh, F. G. Prognosis in schizophrenia: a preliminary report. *Amer. J. Psychiat.*, 1938, **95**, 583–596.

69. Ross, J. R., & Malzberg, B. A review of the results of the pharmacological shock therapy and the metrazol convulsion therapy in New York State. *Amer. J. Psychiat.*, 1939, **96**, 297–316.

70. Rupp, C., & Fletcher, E. K. A five to ten year follow-up study of 641 schizophrenic cases. *Amer. J. Psychiat.*, 1940, **96**, 877–888.

71. Sakel, M., & Gjessing, R. Discussion. In Symposium—treatment of schizophrenia. *J. ment. Sci.*, 1938, **84**, 685–692.

72. Slater, E. T. O. Evaluation of electric convulsion therapy as compared with conservative methods of treatment in depressive states. *J. ment. Sci.*, 1951, **97**, 567–569.

73. Smith, A. Planned research. *Trans. Amer. Ther. Soc.*, 1948, **49**, 73–75.

74. Smith, L. H., Hastings, D. W., & Hughes, J. Immediate and follow-up results of electro-shock therapy. *Amer. J. Psychiat.*, 1943, **100**, 351–354.

75. Stengel, E. A follow-up investigation of 330 cases treated by prefrontal leucotomy. *J. ment. Sci.*, 1950, **96**, 633–662.

76. Stunkard, A. J. A method of evaluating a therapeutic agent. Results in a study of dibenamine. *Amer. J. Psychiat.*, 1950, **107**, 463–467.

77. Tait, C. D., & Burns, G. C. Involutional illnesses: a survey of 379 patients, including follow-up study of 114. *Amer. J. Psychiat.*, 1951, **108**, 27–36.

78. Taylor, J. A., & Van Salzen, C. V. Prognosis in dementia praecox. *Psychiat. Quart.*, 1938, **12**, 576–582.

79. Tillotson, K. J., & Sulzbach, W. A comparative study and evaluation of electroshock therapy in depressive states. *Amer. J. Psychiat.*, 1945, **101**, 455–459.

80. Whitehead, D. Improvement and recovery rates in dementia praecox without insulin therapy. *Psychiat. Quart.*, 1938, **12**, 409–413.

81. Wilson, D. C. The results of shock therapy in the treatment of affective disorders. *Amer. J. Psychiat.*, 1939, **96**, 673–679.

82. Ziskind, E., Somerfeld-Ziskind, E., & Ziskind, L. Metrazol therapy in the affective psychoses: study of a controlled series of cases. *J. nerv. ment. Dis.*, 1942, **95**, 460–473.

83. Ziskind, E., Somerfeld-Ziskind, E., & Ziskind, L. Metrazol and electric convulsive therapy of the affective psychoses. *Arch. Neurol. Psychiat.*, 1945, **53**, 212–217.

84. Zubin, J. Design for the evaluation of therapy. *Res. Publ. Ass. nerv. ment. Dis.*, 1951, **31**, 10–15.

85. Zubin, J. Evaluation of therapeutic outcome in mental disorders. *J. nerv. ment. Dis.*, 1953, **117**, 95–111.

Received August 16, 1956.

APPENDIX

Key to Tables

Types of Patients

A	Alcoholics
AfP	Affective psychoses MD-d; MD-m; IM
AOP	All other psychoses
AP	All psychoses
C	Carcinoma
CA	Psychosis with cerebral arteriosclerosis
CS	Psychosis with cerebral syphilis
D	Drug psychosis
DP	Dementia praecox
Dp	Depressed states
E	Epilepsy
En	Encephalitis
G	General paresis
Gpl	General paralysis
IM	Involutional melancholia
IM-M	Involutional melancholia, melancholic
IM-Pa	Involutional melancholia, paranoid
M	Mixed psychoses
MD-ag.d.	Manic depressive, agitated depression
MD	Manic depressive psychosis
MD-d	Manic depressive, depressed
Mdef	Mental deficiency
MD-mix.	Manic depressive, mixed
MD-m	Manic depressive, manic
Mo-h	Morphia hallucinosis
N	Neurosis
O	Other
Ob	Obsessive states
OBD	Organic brain damage
OBND	Other brain or nervous diseases
Pa	Paranoid
PN	Psychoneurosis
P.inf.	Psychosis with infection
P.pell.	Psychosis with pellagra
P.som.	Psychosis with somatic disease
PsPath.	Psychopath
RD	Reactive depression
SA	Senile arteriosclerosis
Sc	Schizophrenia
SP	Senile psychosis
U	Unclassified
Ud	Undiagnosed
Uns.	Unspecified
V	Varied

Duration of Illness

Ac	Acute	•	Average
Ch	Chronic	°	About
SAc	Subacute	<	Less than
Us	Unspecified	>	More than
Va	Varying		

Duration of Follow-up

Im	Immediate	•	Average
Us	Unspecified	°	About
Va	Varying	<	Less than
P.Ad.	After admission	>	More than
P. F.Ad.	After first admission		
P. Dis.	After discharge		

Types of Somatotherapy

C & ECT	Cardiazol and electroconvulsive therapy	Me	Metrazol
		ECT	Electroconvulsive
I	Insulin	L	Lobotomy
I•	Insulin with few comas	PNa	Prolonged Narcosis

Results

D	Dead	(a)	Paroled successfully
I	Improved	(b)	Complete & social recovery
MI	Much improved	(c)	Improved & slightly improved
R	Recovered	(d)	Full recovery
U	Unimproved		
OD	Otherwise discharged		

Psychological Bulletin
Vol. 54, No. 3, 1957

A MICROGENETIC APPROACH TO PERCEPTION AND THOUGHT

JOHN H. FLAVELL and JURIS DRAGUNS
University of Rochester

It is the purpose of this paper to present a theoretical approach to perception and thought which, although by no means entirely new, will undoubtedly seem strange and unorthodox to many. The term "microgenesis," first coined by Werner (**132**) as an approximate translation of the German word *Aktualgenese*, will refer here to the sequence of events which are assumed to occur in the temporal period between the presentation of a stimulus and the formation of a single, relatively stabilized cognitive response (percept or thought) to this stimulus. More specifically, the term will refer primarily to the prestages of extremely *brief* cognitive acts, e.g., the processes involved in immediately perceiving a simple visual or auditory stimulus, conceptually generating a word association, etc. Thus, cognitive sequences involving many seconds or minutes, such as perceptual changes resulting from prolonged fixation, will not be considered here as examples, or at least as typical examples, of microgenetic development. Within this somewhat restricted conception of microgenesis or microdevelopment, one can distinguish, in terms of experimental operations, between "microgenesis of thought" and "microgenesis of perception." In the former case we refer to situations in which little attention is given to the conditions of stimulus or task presentation but careful attention is paid to the temporal development of the conceptual response. In the latter case, we refer to conditions in which considerable attention is paid to the manner of stimulus presentation but little, if any, is paid to the temporal evolution of the ensuing verbal response. The experimental paradigm of microgenesis of thought consists of presentation of a stimulus to cognition, under optimal conditions of perceptual "intake," and some sort of attempt to study, or even control, the evolution of the cognitive response to this stimulus. The paradigm of microgenesis of perception, on the other hand, usually entails the successive presentation of a stimulus under conditions of increasing clarity. Successive tachistoscopic presentation of visual stimuli with exposure times gradually increasing until complete perception is possible, considered as the experimental homologue of the everyday, near-instantaneous process of simply "seeing" an object, would perhaps be the best example of this paradigm. It should be mentioned that the distinction made here stems from a distinction between typical experimental conditions and does not imply a particular brief for or against any basic dichotomy between perception and thought.

In attempting to conceptualize processes within a microgenetic framework, at least two basic questions arise. From the evidence available, what formal principles of cognitive microdevelopment have been or could be derived to constitute a first beginning of a microgenetic theory? Of what use would such a theory be in organizing known facts of normal and abnormal perception and thought and in constructing testable hypotheses for future research? It is hoped that this paper will suggest partial answers to these questions. We shall first survey some of the theoretical and experimental work which seems to us to bear upon the first of the two questions. Following this, some tentative notions will be proposed with regard to the second question.

Microgenesis of Perception

There is a fairly sizable body of literature concerned, in one way or another, with the temporal evolution of percepts. A good half of these studies emanate directly from one microgenetically oriented "school" and, in this sense, form a tightly knit whole. The remainder of the investigations differ widely among themselves as to theoretical orientation, experimental procedure, etc. In this section we will first describe the contributions of the former group of studies and then compare their findings with those of the miscellaneous remaining experiments.

In the early twenties, there arose in Germany a movement against post-Wundtian elementaristic psychology led by Felix Krueger of Leipzig. Like the better-known Berlin group, Krueger and his followers were Gestaltists and stressed the intrinsic structuredness of perception. Unlike the Berlin school, however, they were particularly concerned with the temporal development of percepts as well as with the formal properties of completed percepts. Krueger developed a complicated and somewhat esoteric general theory which is of only tangential relevance to microgenesis (**59**). His co-worker Sander, however, did develop an explicitly microgenetic theory of perception within Krueger's framework (**94**, **95**) and, with his students, carried out a variety of experimental studies on the problem. He believed that perception is a developmental process consisting of a number of conceptually distinct phases. Further, he assumed that percepts obtained under inadequate stimulus conditions, e.g., brief tachistoscopic exposure, are essentially the same as the initial, transitory percepts which precede the final perceptual response under normal stimulus conditions. He granted that the precursors of the final percept are not observable in the normal, perceptual process. However, he argued that if one experimentally blocks the formation of clear, complete percepts by presenting stimuli very briefly, in bad lighting, in peripheral vision, etc., one can elicit these perceptual precursors. On the basis of experimental findings, Sander was able to offer a fairly detailed description of perceptual microgenesis or *Aktualgenese*, as he called it. Our account of the process will follow that of Undeutsch (**112**), one of Sander's students.

When a perceptual stimulus is presented under conditions of gradually increasing clarity, the initial perception is that of a diffuse, undifferentiated whole. In the next stage figure and ground achieve some measure of differentiation, although the inner contents of the stimulus remain vague and amorphous. Then comes a phase in which contour and inner content achieve some distinctness and a tentative, labile configuration results. Finally, the process of Gestalt formation becomes complete with the addition of elaborations and modifications of the "skeletal Gestalt" (*Gestaltgerüst*) achieved in the previous stage.

As development proceeds, external, objective characteristics more and more supplant inner, personal factors as determinants of the structure perceived. As Undeutsch puts it, the balance of endogenous to exogenous determinants changes as perceptual microdevelopment proceeds. Of particular interest to Sander and his students was the stage just preceding the formation of the final, stable percept. In this *Vorgestalt* or preconfiguration phase the *S* has constructed a tentative, highly labile Gestalt which is more undifferentiated internally, more regular, and more simple in form and content than is the final form which is to follow it. The construction of this initial, flux-like pre-Gestalt is said to be accompanied by decidedly unpleasant feelings of tension and unrest which later subside when a final, stable configuration is achieved. The emotionally-charged character of the *Vorgestalt* stage is stressed by many investigators (**38**, **40**, **65**, **96**, **107**, **116**, **136**) whose reports are often supplemented by colorful and dramatic verbal reports by the *S*s.

These then were the near-unanimous conclusions of Sander and his students with respect to the microgenesis of percepts. Under what experimental situations were these findings obtained? The Sander group showed no lack of imagination in their efforts to study *Aktualgenese* under all possible conditions. Some investigators presented stimuli under gradually increasing tachistoscopic exposure time. Using this technique, paintings by famous artists (**65**), three-dimensional geometric figures (**38**), and groups of everyday objects (**73**) were presented to *S*s and percepts elicited at each exposure time were recorded and analyzed. Sommer (**107**) varied this procedure by gradually decreasing, rather than increasing exposure time and was able to show, in reverse order, the usual sequence of developmental stages. Wohlfahrt (**136**) presented geometric designs in extreme miniature at first and gradually increased their size until *S*s were able to see them clearly and without effort. Butzmann (**9**) also using geometric designs, recorded perceptual alterations as stimuli were gradually moved from the extreme periphery of the visual field in towards a central fixation point. Other investigators used stimuli or arrangements of stimuli which were meaningless or disorganized and compared perceptual development under such conditions with that which occurred when meaningful, organized stimuli were used (**23**, **47**, **48**). Additional investigations conducted by the Sander group involve the microgenesis of tactile impressions (**40**), the temporal process of describing clearly seen objects (**116**), and miscellaneous other problems (**93**, **96**). Such was the variety of stimulus conditions employed. The perceptual responses on which the theory was based were obtained in either of two ways: (*a*) simple introspection, or verbal report (**40**, **65**, **96**, **107**); (*b*) pictorial reproduction of what was perceived (**9**, **23**, **47**, **48**, **73**, **136**), supplemented in one case by manual arrangement of concrete stimulus objects in attempted duplication of the percept (**40**).

It is thus apparent that Sander and his group made a vigorous and concerted attack on what they saw as an important problem in perception. In reading through the variety of parallel studies done outside of Germany one is struck by the fact that the great bulk of *Aktualgenese* research is seldom cited. Similarly, references to non-German experiments on perceptual microgenesis are equally infrequent in the work of the Krueger school. This lack of cross fertilization, however unfortunate in some ways, does make it possible to compare the experimental conclusions of scientists who are not mutually tainted by each other's theoretical preconceptions. It is therefore interesting to note that Sander's assertion that microgenesis begins with diffuse, whole percepts which subsequently become sharpened and internally differentiated receives considerable confirmation from other studies. For example, experimenters using such different stimuli as geometric figures (**6**), letters of the alphabet (**20**), Rorschach (**80**, **109**) or self-made (**31**) inkblots, Rubin figure-ground cards (**134**), and various kinds of pictures (**12**, **22**, **103**), have also reported developmental sequences in the general direction of diffuse to specific. Further, Brigden (**6**) found a tendency towards simplification, completion, transposition, and increased symmetry as development progressed—a finding quite congruent with Sander's statement that percepts at the *Vorgestalt* stage tend to be made "better Gestalten" at the expense of object similarity. Brigden also lists an early tendency to complicate the percept which the Sander group did not explicitly postulate. It is, however, possible that this complication tendency is not unlike the microgenetically early overinvestment of meaning noted by Dyn (**23**), Hippius (**40**), and Johannes (**47**, **48**). As to the microgenetically late trend from specific details to integrated wholes, tachistoscopic studies using Rorschach stimuli confirm this only in part (**80**, **109**). In these latter studies a continuous, nonreversing trend from wholes to details is found, although those whole responses which *are* given in the end stages do tend to be of the integrated, internally differentiated rather than global type. On the debit side, the intense emotionality which the Sander group reports as an invariable concomitant of *Vorgestalt* formation is certainly not stressed by most other investigators, although Douglas (**22**) makes explicit mention of it. There are other minor disagreements between the findings of the *Aktualgenese* group and those of other investigators. Since the experimental methods used in the studies to be compared are often only roughly equivalent, it is difficult to interpret the meaning of such disagreements with any confidence.

Before concluding our account of experimental studies of perceptual microdevelopment it must be mentioned that many of these studies would be considered quite poor by present-day methodological standards. This is especially, although not exclusively, true of the research done by Sander and his school. Few *S*s were used and these were seldom experimentally naive, statistics were inadequate or absent, and methods of measuring and evaluating perceptual responses were informal to say the least. In addition, serious questions concerning basic assumptions can be posed, as will be seen later. Nonetheless, the existing German and non-German studies together constitute a rather extensive and exciting first assault on the truly fundamental problem of how our percepts get formed. As will soon be apparent, there has been considerably less systematic experimental work done on the equally fundamental problem of how our thoughts develop.

Microgenesis of Thought

If one wished to apply the term "microgenesis of thought" to all published accounts of the cognitive steps involved in solving a problem, the relevant literature would be vast indeed. Humphrey (**44**), Johnson (**49**, **50**), Osgood (**75**), Vinacke (**118**, **119**), Woodworth (**137**), Woodworth and Schlosberg (**138**), and others have given ample reviews of the multitude of studies which describe the temporal sequence of concept acquisition or problem solution. Likewise, there are a number of published accounts of the microdevelopment of creative thinking, most of which have been reviewed by Vinacke (**119**) and Woodworth (**137**). Wallas (**122**), for example, divided the development of a creative thought into four stages: preparation, incubation, illumination, and verification. Patrick (**76, 77, 78**) and Eindhoven and Vinacke (**25**) conducted laboratory studies which attempted to test Wallas' assertions. Although to define the limits of a single thought formation is admittedly a hazardous procedure, it may be fairly safe to assume that many, many thought formations occur in any solution sequence as extended as those typically involved in studies of creative thinking, formal problem-solving (**24**), and the like. It may be that laws of cognitive development in a solution process which extends over hours, days, or even years are of a piece with those pertaining to a "single" thought which requires seconds or fractions of a second to run its course, although at present we see no good evidence for such an identity. In any case, the present discussion will be confined primarily to those few studies which concern the nature of thought in relatively brief cognitive sequences.

As is well known, the classical controversy between the Cornell and Würzburg schools involved, among other things, a dispute as to whether images were or were not the "carriers" of mental life (**4, 44**). In their attempts to settle the question by means of introspection studies the members of these schools were of necessity concerned with what lay behind completed cognitive acts. Although they were not explicitly concerned with constructing microgenetic theories, some of their findings bear upon the development of thought as we are defining it. For example, despite differences in opinion as to whether or not thoughts are fundamentally imaginal in substance, both factions reported evidence that images may play a variety of roles in the microgenetic sequence. Thus, according to Humphrey's account (**44**, pp. 283–288), various introspective studies suggest that images sometimes seem merely to illustrate or accompany thoughts already in progress, sometimes serve as starting-points for subsequent thought microgenesis, and sometimes even constitute distractions by leading *S* to dwell upon the images instead of progressing in the thought sequence or by leading *S* to a thought wholly irrelevant to the cognitive task at hand. Further, Willwoll (**44**) found that the images which impede thinking tend to be more clear and concrete than those which do not. Although its function may be highly variable from instance to instance, it is perhaps safe to conclude that imagery, when it occurs, tends to be a phenomenon characteristic of the earlier stages of thought microdevelopment.

In addition to their studies of the role of imagery in the microdevelopmental process, these early psychologists, especially the Würzburgers, made some interesting observations about the developmental sequence as a whole. Thus Messer (**70**) distinguished between vague, undeveloped thoughts without words or images and fully formulated propositions with clear consciousness of meaning. For example, one of his *S*s gave the following introspection after having responded "corner" to the stimulus word "angle": "The tendency was towards the well-known proposition that the sum of the angles of a triangle equals two right angles . . . but it did not mature" (p. 178). Bühler (**8**) studied somewhat more complex thought problems, instructing his *S*s to "solve" a variety of proverbs, aphorisms, etc., and to report their introspections of the solution process. On many occasions the *S*s would report that they had, early in the solution sequence, vague, imageless half-thoughts or premonitions about such things as the task, the nature of the solution, the possibility or impossibility of solution, the problem's relationship to other problems, etc.[1] Bühler's data suggest that very early thoughts seem to serve somewhat as global schemata which orient the thinker as to the nature of the solution. That is, the thinker may have experiences of vaguely knowing where the solution will lie, with what problems or persons the solution is associated, how difficult the solution will be, and so on, considerably prior to possessing the fully formulated solution—prior to thinking the problem through. The *S*'s introspections suggest that it is as if the final solution somehow differentiates out of the diffuse, generic-like, "framework" thoughts which precede it. Whether or not these microdevelopmentally immature thoughts always have an image-like composition is a question which seems less important to us today than does the question of the role these early thoughts play in the developmental process.

There have been a few psychologists, from the beginning of the century down to the present day, who have more or less explicitly theorized about the microgenesis of thought. One of the earliest of these was Jung (**51**), who considered the problem in the context of his studies of word association. He expressed the belief that "superficial" word association responses, such as clangs and word and phrase completions, are the initial, immediate cognitive responses to words and that they are normally suppressed in favor of the more meaningful responses which follow them in the apperceptive process.

Jung posited a temporal hierarchy of modes of word cognition which progresses, in the course of the apperceptive process, from the most superficial cognition of the physical characteristics of the word, through a cognition of the word as a member of a familiar phrase, and finally through cognition of the word's denotative and connotative meanings.

Somewhat later, Pick and Thiele (**81**), Van Woerkom (**113, 114**), and Bouman and Grünbaum (**5**) formulated hypotheses about thought development in the course of their work with aphasics. Pick and Thiele, drawing upon the earlier work of Bühler and Messer, suggested that the word cognition process typically goes through a series of stages which are, in part, somewhat reminiscent of Jung's formulation: (*a*) recognition of the word as a physical object, an "acoustic Gestalt," (*b*) an awareness of the general "meaning sphere" of the word, i.e., location of the word in conceptual space, (*c*) comprehension of the grammatical form of the word. Pick and Thiele state that the succession of these stages is not invariable and that more stages may be involved if *S* is required to make a verbal formulation of his cognition. Bouman and Grünbaum conceive of the

[1] For a wealth of anecdotal evidence for such microdevelopmentally early thoughts, see Wallas (**122**, Chap. 4).

cognition of a stimulus as beginning with a total, amorphous general impression (e.g., "good" or "bad," "right" or "wrong") which, in normal individuals, is followed by successive differentiations of the total stimulus into its component meaningful parts. Similarly, Van Woerkom insists that the developmental process typically begins with the conception of the whole idea, with a stage of analysis and synthesis following.

A more recent theoretical exposition of the course of thought development in forming word associations is given by Rapaport, Gill, and Schafer (**84**) and Schafer (**97**). They suggest that the normal process of giving a word association to a stimulus word consists of two principal microdevelopmental phases: an analytic, decompository stage in which the stimulus word is broken down into its component ideas and one of these ideas is selected as the basis for the association to come; following this, a synthetic, compository phase in which the response word is constructed from a thought associated with this particular component idea. In both phases the associative process is assumed to be guided by an over-all set to produce a response word conceptually related to the stimulus word, a set which becomes even more specific when *S* hears the stimulus word. When for any reason the thought process does not pass through both phases the resulting associative response will be atypical. Rapaport et al. designate as *close* those responses which indicate that the process has not proceeded past the first, analytic phase and as *distant* those which suggest that the synthetic process has overdeveloped in an associative sequence tangential or irrelevant to the task-induced anticipation. Thus, *close* associations include repetitions of the stimulus word, attributes, clangs, and phrase completions—associations which indicate, as Jung had suggested earlier, that the associative process has been "aborted" early in its microdevelopment. Those responses which are logically unrelated or very marginally related to their stimulus words are scored as *distant* associations, the presupposition being that intermediary associations in the synthetic phase have constituted the connecting links between the stimulus word and the seemingly irrelevant response word. Despite differences in basic theoretical orientation, Jung and the Rapaport group appear to agree, at least implicitly, on several points of importance to microgenetic theory. First, they both consider the task of giving a word association to a verbal stimulus as a simple thought problem, the study of which may shed light on cognitive processes in general. Further, they believe that producing word associations is a microdevelopmental process in which successive, and perhaps conceptually distinct stages occur within a brief time span.

By far the most explicit theoretical elaboration of a microgenetic view of thought formation has been given by Schilder (**98**, **99**). According to this theorist, thought begins with a diffuse conception of its goal, some sort of vague direction in which it is to go. The early stages of its development from this point onward he termed the ***preparatory phase of thought***. In this phase a host of mental contents (***presentations*** as Schilder called them) feed into the ongoing thought development. These vague ***presentations*** may be logically relevant or irrelevant in relation to the thought nucleus which is at this time gaining ever-increasing structure and clarity. Those ideas or images which are relevant are incorporated into the process and enrich the forming thought; those which are irrelevant normally get suppressed and at most remain only as "background music" for the evolving thought. In this early, preparatory period mental contents are said to be of a symbol- and imagelike character, very susceptible to fusions and condensations with each other and with the developing thought structure, and subject to emotional restructuring in accordance with what we would today term "primary process" influence. The logic by which certain of these primitive presentations rather than others come to the fore is not specifically described beyond stating that contiguity and similarity, especially similarity of external, superficial attributes, play major roles. Schilder further states that, as the development progresses, the thought structure normally becomes more and more reality-oriented and less and less wish-determined—Undeutsch (**112**) had said the same thing about perceptual development—as well as less ridden with concrete imagery, less symbolistic, less undifferentiated and unstable, etc. Schilder's theory of microgenesis can of course be roundly criticized on a number of grounds. The referents of many of his terms are highly obscure, his exposition proceeds unencumbered by restraint or caution, his thesis lacks direct evidence, and so forth. Nevertheless, it can be said that he has fashioned a series of strikingly imaginative and original hypotheses about an aspect of cognition which has sadly needed explicit theorizing, however high-flown and speculative.

At this point our rather meager history is completed and stock-taking is in order. Although the existing evidence hardly permits any kind of integrated theory of thought microdevelopment, it is at least possible to see some commonality and consistency in what has been said and to organize a series of very tentative statements about the topic—a sort of loose conceptual framework within which to think about the development of thoughts. In this hypothetical account, we will lean most heavily upon Schilder's writings but will also draw from the work of Rapaport et al., Jung, and the rest.

First of all, thought in its early stages is global, diffuse, and undifferentiated in structure (**131**); that is, mental contents, be they images or imageless thoughts, tend to coexist without articulation and without clearly defined interrelationships. These early thought elements may be vague, imageless thought tendencies concerning the task, the solution, the thinker's relationship to task and solution, etc. Images, when they occur in thinking, also tend to be early rather than late products and may serve as primitive and concrete anchoring-points or, as Schilder puts it (**99**), "symbols" for what is to come. Microgenetically early thoughts, imaginal or imageless, seem to have the quality of what Rapaport (**83**) has termed *drive-representations*, i.e., needs and affects are particularly sovereign in determining which thoughts push for expression, which thoughts feed into the developmental process. Moreover, the laws of combination and association of thoughts in the beginning phases likewise seem to resemble those posited for primary process thinking, i.e., association by contiguity, association by superficial, external similarity, association on the basis of common personal predicates and a prevalence of condensation and displacements (**32**). Thus the thought process tends first towards this, then that premature, "paleological" solution (**1**) and early judgments of solutions tend to be primitive, dichotomous affairs framed in terms of me-not me, good-bad, etc. (**99**). In the later stages of development thought ordinarily becomes differentiated into various components and these components become logically interrelated in the formation of the solution. Thought in the final phase is normally reality- rather than drive-oriented and the early nonlogical thought developments have become aborted, as it were, and no longer influence the form of solution. It is very likely that, in most people under ordinary circumstances, this extraordinarily rapid developmental process does not become an object of awareness and the thinker is conscious only of the completed thought.[2]

Implications of Microgenetic Theory

It is proposed here that the microgenetic approach can be fruitfully applied to the cognition (perception and thought) of pathological individuals under normal conditions and of normal individuals under atypical, nonnormal conditions. An attempt will be made to provide evidence that such atypical cognitions tend to manifest formal characteristics similar to those already predicated for microgenetically incomplete cognition. Such evidence would suggest the general hypothesis that most or all atypical cognitions, whether found in normal or pathological individuals, are special cases of normal, mature cognition in the sense that they are cognitive forms which have aborted prior to complete development. Thus, within this frame of reference, normal cognition is not defined simply by the absence of nonnormal attributes nor is atypical cognition viewed as a unique, qualitatively distinct formation. Normal, logical cognition is seen as a microdevelopmental achievement of the organism and deviations therefrom as developmental arrests. Such an approach, should the facts justify it, permits one to subsume a host of cognitive phenomena under one developmental theory and, at the same time, makes the study of the normal, prototypical microgenetic process something of considerable theoretical urgency.

In surveying the evidence for these beliefs, our previous major breakdown in terms of percepts versus thoughts will be abandoned; instead, we shall examine the findings topic by topic, drawing from whichever set of microgenetic hypotheses (perceptual or thought) best applies to the data at hand.

Normals Under Atypical Conditions

Distraction constitutes one set of conditions under which normal individuals tend to produce cognitive responses which could be called atypical. There have been a few studies which have attempted to study distraction effects. Jung (**51**) and Speich (**108**), for example, both found that when *S*s are asked to give word associations under distraction conditions the tendency is for superficial, external responses (clangs, word-completions, etc.) to increase. In another publication (**52**) Jung re-

[2] In this connection Rapaport, Gill, and Schafer (**84**) state:

"These preparatory phases are, in the average subject, preconscious: however, in introspective and/or obsessive people, the inquiring examiner often obtains reports on what happened in the brief interval between the stimulus- and reaction-word—how definitions, images, clang and other deviant associations occurred and were rejected, though the result came quickly and as a 'popular reaction' " (p. 20).

ports an interesting early study by Stransky on the effects of an experimental condition similar to distraction. His *S*s were instructed to talk about anything for one minute without attending to what they were saying. He found that these instructions produced an abundance of immature-like processes which included substitution of superficial connections (clangs, etc.) for logical ones, numerous perseverations, and fusions of competing verbal responses which resulted in neologisms and contaminations. Not all the evidence with regard to distraction effects is in accord with microgenetic theory, however; Cameron and Magaret (**11**) failed to find such effects when distraction was superimposed on a task of completing incomplete sentences.

There is some evidence pertaining to the formal characteristics of thinking in dreams, daydreams, and semisleep. Freud (**32**), as is well known, characterized dream-thinking as being replete with condensations, displacements, symbolization of abstract thoughts via concrete images, prelogical thinking mediated by external and superficial or highly subjective similarity, etc. Varendonck (**115**), in his classical study of daydreams, has likewise stressed the lack of criticality and logical direction and the important role of nonverbal imagery which obtains in ordinary, conscious fantasy. Mintz (**72**), Rapaport (**82**), and Silberer (**102**) have described the hypnagogic or semisleep state in somewhat similar terms: decrease in reflective awareness, or sharply focussed self-criticality; symbolization (via images rather than words) of bodily states, attitudes, etc., as well as ordinary thought contents; and a tendency to substitute prelogical autistic thinking for logical, conventional thought. Jung (**51**) reports a study in which one *S* was given a word-association test both under normal waking conditions and under conditions of semisleep. The *S*, while drowsy, gave about seven times as many clang reactions as when in the waking state.

There are a variety of studies describing the effects of various drugs upon thought and perception. Smith (**106**) found that alcohol tends to increase the frequency of word associations of Jung's "outer" type. He did not report his results statistically but the senior author's recalculations of Smith's data suggest that this tendency was significant at about the $p<.15$ level of confidence. Both Woodworth and Schlosberg (**138**) and Kohs (**57**) allude to old studies by Kraepelin and his students which suggest that caffeine tends to cause *S*s to give more superficial word associations. There are a number of studies describing the effects of mescaline, lysergic acid derivatives (LSD), and other "psychotogenic" drugs on cognition (**28**, **37**, **41**, **42**, **43**, **45**, **63**, **64**, **69**, **92**, **110**). Some, although by no means all, of these drug effects seem consistent with what we would consider to be the formal characteristics of microdevelopmentally early cognition. Thus *S*s under the influence of LSD or mescaline have been found to show, among other things: looseness of association; rhyming and punning; inability to follow a single train of thought without interpenetration and fusions with other thought sequences; predominance of vivid imagery in thinking; and a general lability of percepts. In connection with the imagelike character of thoughts, for example, some of Meadow's *S*s reported that they had to overcome the ever-present visual images in order to think abstractly (**69**). One of Guttman's *S*s described this phenomenon as follows: "Each word I thought was connected with a picture. This hindered my thinking, as the concrete pictures held me" (**37**, p. 213). Lindemann and Clarke (**63**), and Kubie and Margolin (**60**) have also suggested that other drugs, such as scopolamine, sodium amytal, nitrous oxide, and various barbiturates, produce cognitive states essentially equivalent to those previously described for the semisleep state.

In addition to distrzction, drugs, deviations from the waking state, etc., there are several other miscellaneous conditions which deserve brief mention. According to Kohs (**57**), Aschaffenburg found the familiar increase in clang and completion responses when *S*s were in a fatigued state. Bexton, Heron, and Scott (**3**) found that prolonged insulation of *S*s from external stimuli caused an increase in directionless thought of the daydream type and a falling back upon extremely vivid imagery. Kline and Schneck (**56**) found more "associative alterations" when *S*s gave word associations while under hypnosis. Although it is not altogether clear from their paper, "associative alteration" appears to include Rapaport, Gill, and Schafer's (**84**) *distant* and *mildly distant* categories primarily. Finally, Gellhorn and Kraines (**34**, **35**), in another word-association study, report that experimentally induced anoxia causes an increase in perseverations and unusual, irrelevant associations—a result consistent with McFarland's earlier findings on the psychological effects of oxygen deprivation (**67**).

We have so far considered verbal cognitive behavior in normals under atypical organismic states. Also of interest from a microgenetic standpoint are subverbal or preverbal cognitive responses under more or less typical organismic states—namely, the kinds of cognitive responses found in studies of semantic conditioning (**21**, **85**, **86**, **87**, **88**, **89**, **90**, **91**, **139**) and subception (**62**, **66**, **68**, **121**). In both semantic conditioning and subception studies *S*s evidence some sort of "cognition" of stimuli (usually below the level of verbal report) by means of measurable electrodermal or salivary responses. It is interesting to speculate as to whether these kinds of dim cognitions which "register" only at the physiological level can be considered microgenetically early, primitive forms which do not, for one reason or another, attain conscious awareness. Some of the studies of semantic conditioning reveal curious facts which might suggest this. Razran (**90**), for example, reports one experiment in which a salivary response was conditioned to a given word and then *S* was presented with a variety of other words, each bearing a different relationship to the original stimulus word. As might be expected, synonyms, supraordinates and contrasts of the original word elicited salivary responses of fairly large magnitude. What was surprising, however, was that ***homophones*** (i.e., clangs) of the original word elicited salivary responses about as large as the more logically respectable coordinates, part-wholes, whole-parts, and predicates, and of greater magnitude than subordinates, a highly logical category! Common sense will assert, and Flavell (**30**) has demonstrated experimentally, that normals do not consciously consider words related by sound similarity to be as similar in meaning as those related in terms of any of the semantic categories mentioned above. Yet the various studies of semantic conditioning seem to indicate that we do make generalizations about verbal symbols, at the physiological level, on the basis of physical as well as semantic similarity. It may indeed be, as Jung long ago suggested, that "every apperceptive process of an acoustic stimulus begins at the stage of clang-like apprehension," and that such an apprehension somehow gets "recorded" in an immediate autonomic reaction but normally does not persist, in subsequent microdevelopment, as a conscious component of the final cognition. It may also be possible to view at least some aspects of the problem of subception in similar terms. Bruner and Postman (**7**) some time ago offered an interpretation of perceptual defense and subception data in terms of levels of response. They suggest that generic and diffuse affective responses may occur prior to, or at lower thresholds than, conscious cognitive responses pertaining to the specific nature of the stimulus. They also mention, in passing, the relationship of this view to the classical "stages of perception" theories we have already reviewed. More recently, Lazarus (**61**) has offered a somewhat similar view as one possible explanation of subception. He suggests that the autonomic nervous system may be capable of making global, all-or-none discriminations between "danger" and "no danger," "shock" and "no shock," etc., under stimulus conditions which are not adequate for precise differentiation of the more complex attributes of the stimulus.

The really intriguing question which all this poses, of course, is whether or not so-called "unconscious" thinking and perceiving can be meaningfully framed within microgenetic theory. One wonders whether the similarities which may exist between unconscious cognition, as in dreams for example, and what we have termed microgenetically early cognition are merely coincidental. May it be that unconscious, primary process cognitions are those which begin to develop, make their mark on behavior, and then, for reasons which can only be guessed at, abort below the level of conscious awareness? Conrad (**15**), to whose work on microgenesis we shall shortly refer, proposed a very similar explanation. The recent experiments by Smith and Henriksson (**104**) and Klein et al. (**55**) also provide some support for such a conceptualization. In these studies it was demonstrated that stimuli flashed at tachistoscopic exposure times too brief for conscious recognition definitely modified the perception of other suprathreshold stimuli presented immediately after them. Klein (**54**, p. 23) makes one statement, in discussing his results, which well expresses the tenor of our own musings:

> A working hypothesis in this situation is that the A figure, exposed for a few microseconds, starts a cognitive process which is interrupted or covered over so quickly by the B process that it is, in effect, aborted. Some kind of compromise formation results in the reported percept. *Such incomplete formations may provide the condition for the operation of primary process mechanisms* (italics ours).

Pathological Individuals

We shall confine our discussion in this section mainly to two diagnostic groups in which atypical cognition seems especially predominant—schiophrenia and aphasia. Fairly adequate accounts of theory and research on schizophrenic cognition may be found in Arieti (**1**), Bellak (**2**), Cameron (**10**), Fenichel (**27**), Flavell (**30**), Goodstein (**36**), Kasanin (**53**), Wegrocki (**123**), and White (**135**). A study of the literature on schizophrenic thought and perception reveals two facts of particular relevance to a microgenetic approach. The first of these pertains to what seems to be a rather striking similarity between microgenetically immature cognition and schizophrenic cognition. The senior author, for example, has elsewhere (**30**) summarized some of the alleged salient features of schizophrenic thinking roughly as follows: condensations, ellipses, word salad, neologisms, clang associations, tangentiality, incoherence, word magic, "paleological" thinking based upon logically superficial predicates of either external or inner-personal origin, excessive use of concrete symbolism, and others. Secondly, the so-called "regression" theorists, i.e., Arieti (**1**), Von Domarus (**120**), Storch (**111**), Vigotsky (**117**), Werner (**131**), White (**135**) and various psychoanalysts (**27**), have related schizophrenic cognition to that found in normals under abnormal conditions, in children, and in people belonging to "less advanced" cultures. That is, they regard schizophrenic cognition as one instance of a more generic, *primitive* mode of cognition which is found in a variety of individuals under various conditions (**30**). Only Schilder (**98**, **99**, **100**) seemingly, has both stressed the formal similarities among primitive or regressive cognitive processes of various kinds and also explicitly taken the further step of viewing such processes as *themselves* possible instances of microgenetically immature cognition. It is of interest to note that Schilder focused particular attention on schizophrenia as the example par excellence of a condition in which early cognitive formations intrude into consciousness and get expressed as though they were completed thoughts. Two of the most interesting recent studies pertinent to the problem of microgenesis in schizophrenia are reported in the article by Phillips and Framo mentioned above (**80**). Rorschach cards were successively shown to schizophrenics and normals under increasing tachistoscopic exposure times and responses scored on a scale of amorphousness-specificity-differentiation and organization, devised by Friedman (**33**). As exposure time increased, the normals' percepts tended to progress from an initial amorphousness and vagueness to specificity and integration; the schizophrenics' percepts, on the other hand, tended to remain at the initial undifferentiated level. Also worthy of mention are the word-association studies by Rapaport et al. cited earlier. They found that schizophrenics exceeded normals in responses presumably indicative of an incomplete associative development, i.e., the various reactions classified as *close* or *distant*.

We have stated that Schilder is essentially the only theorist who has systematically described schizophrenic cognition in microdevelopmental terms. In this respect aphasia has fared somewhat better. As mentioned earlier, Bouman and Grünbaum, Pick and Thiele, and Van Woerkom derived their conceptions of normal microgenesis directly from studies of thought and perception in aphasia. Thus Bouman and Grünbaum, for example, found that an aphasic patient had difficulty in coping with stimuli which required analyzing parts within a whole (e.g., a design embedded within two overlapping figures) but could adequately handle perceptual and conceptual situations in which only a diffuse, over-all apprehension or a total dichotomous judgment was required. From evidence of this kind Bouman and Grünbaum, Van Woerkom, etc., drew two conclusions: first, the normal sequence of cognition has a certain characterizable developmental form; second, this developmental sequence somehow gets arrested during its early stages in aphasia. Certainly the most vocal and explicit proponent of a microgenetic interpretation of aphasic cognition has been Conrad (**13**, **14**, **15**, **16**, **17**, **18**). Conrad has systematically applied the theoretical formulations of the *Aktualgenese* school to aphasic cognitions. He states that the normal process of cognition involves both progressive differentiation and integration of stimulus material and that in aphasia, one or both of these processes typically tends to be incomplete (**14**). Conrad describes four levels of disability which may occur: (*a*) normal Gestalt formation gets accomplished but with abnormal effort and tension; (*b*) the figure gets differentiated from background but does not itself become differentiated or structured; (*c*) figure and ground are not clearly articulated and the percept is vague and amorphous, as though presented tachistoscopically at very brief exposure times; (*d*) lack of any Gestalt formation of any kind (**13**). Conrad suggests that certain memory processes may also be considered from a microgenetic standpoint (**15**). For example, he elaborates upon Wenzl's (**126**) earlier account of the process of word-finding, suggesting that, in attempting to remember a forgotten word, we pass through successive stages structurally similar to those found in perceptual microdevelopment. The sequence of mnemonic reconstitution is the same for normals and aphasics, although, of course, the problem of searching one's memory for forgotten words may be an almost ever-present one for the aphasic. Worthy of citation here also is the extensive work of Ombredane, whose findings likewise support a microgenetic conception of aphasic disorders (**74**). Werner's paper (**132**), a revised and extended version of an earlier German publication (**129**), is the most recent exposition of an avowedly microdevelopmental approach to aphasia, and perhaps the only one in the literature published in English. In this study, one of a series of pioneer investigations in the area of microdevelopment (**127**, **128**, **130**), words were presented repeatedly under gradually increasing tachistoscopic exposure times and normal *S*s were asked to recount their perceptual experiences at each exposure until full recognition was achieved. Werner found that a number of his *S*s reported experiencing spheres of meaning prior to specific and complete recognition of the word stimuli. For example, some *S*s would experience "feelings" about the as yet undiscriminated stimulus word—feelings that it is "warm," "vibrating," "soft," etc. Also, *S*s would occasionally get a global impression of the domain or class within which the word belongs ("it is something shining," etc.). Werner then describes highly similar spheric experiences reported by aphasics in the course of attempting to name familiar objects, read or grasp the meaning of familiar words, and so on. He suggests that in such cases the patient's overt response is the result of a premature precipitation of spheric experiences into verbal expression, i.e., a microgenetic abortion of the kind Conrad and others have described. In the remainder of his paper, Werner discusses some interesting implications of this view for the re-education of patients with aphasic disorders.

Future Problems

We have discussed some of the evidence pertaining to a microgenetic approach to cognition and some of the possible applications of this approach to various cognitive states in normal and pathological individuals. Other possible extensions could be delineated. For example, formal relationships between microgenetically immature cognition and cognitive functioning in nonaphasic brain-damaged cases, aments, depressives, manics, and normal children have not been discussed, although there is some evidence which might support some such comparisons (**26**, **39**, **46**, **79**, **84**, **133**). Also, possible relationships between personality variables and microgenetic sequences need exploration. It is interesting to note that members of the *Aktualgenese* school were actively concerned with correlating individual differences in microdevelopmental sequence with "personality types" (**23**, **38**, **40**) and that Sander himself thought of microgenesis as a potential avenue for the exploration of the unconscious (**94**, **95**). A recent study by Smith and Klein (**105**), although concerned with somewhat more extended cognitive sequences than those we have been considering, is also relevant to the problem of microgenesis-personality relationships. However, the most extensive and perhaps the most intriguing investigations in this area are those recently described by Kragh (**58**). This Swedish psychologist has formulated a bold and explicit personality-perceptual microgenesis theory and has reported a series of tachistoscopic experiments which purport to show relationships between the ontogenesis of personality and the microgenesis of percepts. As a final extension, one can speculate with Werner (**132**) as to whether such .functions as memory and motor performance—as well as perceptual and conceptual developments more complex and of longer duration than the ones considered in this paper—typically undergo developmental sequences similar in formal aspects to those already described.

Such questions, however, seem somewhat premature at present in that they assume a more complete factual knowledge of the prototypical microgenetic processes than we now possess. A problem of much higher priority concerns whether, and by what means, the nature of these elusive processes themselves can be experimentally elucidated. With regard to perceptual microdevelopment, it is clear that more adequately designed studies of the formal aspects of genetic sequences are needed. For example, it would be possible to avoid the hazards of relying solely upon verbal report in tachistoscopic studies by requiring artistically trained *S*s to draw rather than describe their percepts at each exposure time level. It should then be possible to study sequences by having the drawings categorized by judges as to such formal features as diffuseness, degree of figure-ground articulation, etc. Such a study would perhaps lay claim to greater objectivity than those hitherto reported. Likewise, for the development of thoughts or concepts, a plausible experimental technique might be that of motivating *S*s to produce word associations under extreme time pressure and comparing the formal aspects of the resultant associations with those produced by *S*s who had not responded under pressure. Techniques of this type have been used with some success in the past (**19**, **51**, **71**, **101**, **108**). Both of the above methods, or modifications thereof, could of course also be used with nonnormal populations in order to study regressive cognition within a microdevelopmental framework.

In concluding, it is perhaps appropriate to underscore the considerable problems which confront the microgenetic approach in its current form. In the first place, the abstractness, looseness of logical structure, and general semantic imprecision which characterizes present-day microgenetic theory may be in part responsible for the ease with which it seems to subsume so many diverse cognitive phenomena. Such a criticism implies that as the conciseness and testability of the theory increases, nature will seem less cooperative and problems of generalization will arise. Likewise, at the data level, it must be apparent that the findings on the basis of which microgenetic hypotheses have been constructed are by no means gilt-edged. For example, many of the studies cited stem from an era when careful experimental control could hardly be called the rule. Perhaps a more serious criticism pertains to the nature of the typical experimental operations by which microgenesis is allegedly demonstrated. It could be argued, for instance, that the fact that an *S* might, under time pressure, produce responses classified within the theory as microgenetically undeveloped does not prove conclusively that such responses really "occur" but are suppressed in the normal, unhurried associative process. It is certainly possible to pose alternative explanations in terms of variations in set or alterations in verbal habit-family hierarchies induced by time pressure. Similarly, there is no absolute proof that the sequence of percepts found when the tachistoscopic method is used is a faithful reflection of the natural process of percept development. Pertinent criticisms of this order have been raised by Weinschenk (**124**, **125**), and Klein.[3] It is true that one can counter such objections with logical arguments and by citing introspective evidence, such as the verbal reports of Rapaport's obsessional group mentioned earlier (Footnote 2). Nonetheless, such objections have real force and the ***experimentum crucis*** which would settle the matter is difficult to conceive at present. For us the microgenetic interpretation has led to a fresh, albeit highly speculative, view of a variety of cognitive phenomena and has suggested certain lines along which research might proceed. We are thus inclined to tolerate its ambiguities for a time out of sheer curiosity to see what will come of it in the future.

Summary

The present paper has proposed a microgenetic approach to perception and thought. Within this approach, thoughts and percepts are believed to undergo a very brief, but theoretically important, microdevelopment. Evidence was offered both to support the possibility that such microdevelopments do occur in the normal process of thinking and perceiving and to suggest some of the formal characteristics of such evolutions. Further, an attempt was made to delineate some of the possible implications of this approach for cognitive functioning in abnormal

[3] Klein, G. S. Personal communication, May 28, 1956.

individuals and in normal individuals under atypical conditions. Finally, consideration was given to current problems and future research possibilities in relation to a microgenetic framework.

REFERENCES

1. Arieti, S. *Interpretation of schizophrenia.* New York: Brunner, 1955.
2. Bellak, L. *Dementia praecox.* New York: Grune & Stratton, 1948.
3. Bexton, W. H., Heron, W., & Scott, F. H. Effects of decreased variation on the sensory environment. *Canad. J. Psychol.*, 1954, **8**, 70–76.
4. Boring, E. G. *A history of experimental psychology.* (2nd Ed.) New York: Appleton-Century-Crofts, 1950.
5. Bouman, L., & Grünbaum, A. A. Experimentell-psychologische Untersuchungen zur Aphasie und Paraphasie. *Z. ges. Neurol. Psychiat.*, 1925, **96**, 481–538.
6. Brigden, R. L. A tachistoscopic study of the differentiation of perception. *Psychol. Monogr.*, 1933, **44**, No. 1 (Whole No. 197).
7. Bruner, J. S., & Postman, L. Perception, cognition, and behavior. *J. Pers.*, 1949, **18**, 14–31.
8. Bühler, K. On thought connections. In D. Rapaport (Ed.), *Organization and pathology of thought.* New York: Columbia Univer. Press, 1951. Pp. 39–57.
9. Butzmann, K. Aktualgenese im indirekten Sehen. *Arch. ges. Psychol.*, 1940, **106**, 137–193.
10. Cameron, N. The functional psychoses. In J. McV. Hunt (Ed.) *Personality and the behavior disorders.* Vol. 2. New York: Ronald, 1944. Pp. 861–921.
11. Cameron, N., & Magaret, A. Experimental studies in thinking: I. Scattered speech in the responses of normal subjects to incomplete sentences. *J. exp. Psychol.*, 1949, **39**, 617–627.
12. Carl, H. Versuche über tachistoscopisches Bilderkennen. *Z. Psychol.*, 1933, **129**, 1–42.
13. Conrad, K. Über den Begriff der Vorgestalt und seine Bedeutung für die Hirnpathologie. *Nervenarzt*, 1947, **18**, 289–293.
14. Conrad, K. Über differentiale und integrale Gestaltfunktion und den Begriff der Protopathie. *Nervenarzt*, 1948, **19**, 315–323.
15. Conrad, K. Das Problem der gestörten Wortfindung in gestalttheoretischer Betrachtung. *Schweiz. Arch. Neurol. Psychiat.*, 1949, **63**, 141–192.
16. Conrad, K. Über das Prinzip der Vorgestaltung in der Hirnpathologie. *Dtsch. Z. Nervenheilk.*, 1950, **164**, 66–70.
17. Conrad, K. Über den Begriff der Vorgestalt Bemerkungen zu dem Aufsatz von Weinschenk. *Nervenartz*, 1950, **21**, 58–63.
18. Conrad, K. Über das Prinzip der Vorgestaltung Erwiderung auf die vorstehende Arbeit von Weinschenk. *Schweiz. Arch. Neurol. Psychiat.*, 1951, **67**, 119–125.
19. Cordes, G. Experimentelle Untersuchungen über Associationen. *Philos. Stud.*, 1901, **17**, 30–77.
20. Dickinson, C. A. The course of experience. *Amer. J. Psychol.*, 1926, **37**, 330–344.
21. Diven, K. Certain determinants in the conditioning of anxiety reactions. *J. Psychol.*, 1937, **3**, 291–308.
22. Douglas, A. G. A tachistoscopic study of the order of emergence in the process of perception. *Psychol. Monogr.*, 1947, **61**, No. 6 (Whole No. 287).
23. Dun, F. T. Aktualgenetische Untersuchungen des Auffassungvorgang chinesischer Schriftzeichen. *Arch. ges. Psychol.*, 1939, **104**, 131–174.
24. Duncker, K. On problem-solving. (Lynne S. Lees, Trans.) *Psychol. Monogr.*, 1945, **58**, No. 5 (Whole No. 270).
25. Eindhoven, J., & Vinacke, W. E. Creative processes in painting. *J. gen. Psychol.*, 1952, **47**, 139–164.
26. Feifel, H. An analysis of the word definition errors of children. *J. Psychol.*, 1952, **33**, 65–77.
27. Fenichel, O. *The psychoanalytical theory of neurosis.* New York: Norton, 1945.
28. Fischer, R. Factors involved in drug-produced model psychoses. *J. ment. Sci.*, 1954, **100**, 623–632.
29. Flavell, J. H. Thought, communication and social integration in schizophrenia: an experimental and theoretical study. Unpublished doctor's dissertation, Clark Univer., 1954.
30. Flavell, J. H. Abstract thinking and social behavior in schizophrenia. *J.*

abnorm. soc. Psychol., 1956, **52**, 208–211.

31. Freeman, G. I. An experimental study of the perception of objects. *J. exp. Psychol.*, 1929, **12**, 341–358.
32. Freud, S. The interpretation of dreams. In A. A. Brill (Ed.) *The basic writings of Sigmund Freud.* New York: Modern Library, 1938. Pp. 181–552.
33. Friedman, H. Perceptual recognition in schizophrenia: An hypothesis suggested by the use of the Rorschach Test. *J. genet. Psychol.*, 1952, **81**, 63–98.
34. Gellhorn, E., & Kraines, S. H. The influence of hyperpnea and of variations in the O_2 and CO_2 tension of the inspired air on word-association. *Science*, 1936, **83**, 266–267.
35. Gellhorn, E., & Kraines, S. H. Word associations as affected by deficient oxygen, excess of carbon dioxide and hyperpnea. *Arch. Neurol. Psychiat.*, 1937, **38**, 491–504.
36. Goodstein, L. D. The language of schizophrenia. *J. gen. Psychol.*, 1951, **45**, 95–104.
37. Guttman, E. Artificial psychoses produced by mescaline. *J. ment. Sci.*, 1936, **82**, 203–221.
38. Hausmann, G. Zur Aktualgenese räumlicher Gestalten. *Arch. ges. Psychol.*, 1935, **93**, 289–334.
39. Hemmendinger, L. A genetic study of structural aspects of perception as reflected in Rorschach responses. Unpublished doctor's dissertation, Clark Univer., 1951.
40. Hippius, R. Erkennendes Tasten als Wahrnehmung und als Erkenntnisvorgang. *Neue Psychol. Stud.*, 1934, **10**, 1–163.
41. Hoch, P. H. Experimentally produced psychoses. *Amer. J. Psychiat.*, 1951, **107**, 607–611.
42. Hoch, P. H., Cattell, J. P., & Pennes, H. H. Effects of mescaline and lysergic acid (d-LSD-25). *Amer. J. Psychiat.*, 1952, **108**, 579–584.
43. Hoch, P. H., Cattell, J. P., & Pennes, H. H. Effects of drugs; Theoretical considerations from a psychological viewpoint. *Amer. J. Psychiat.*, 1952, **108**, 585–589.
44. Humphrey, G. *Thinking: an introduction to its experimental psychology.* New York: Wiley, 1951.
45. Hyde, R. W., Von Mering, O., & Morimoto, K. Hostility in the lysergic psychosis. *J. nerv. ment. Dis.*, 1953, **118**, 266–267. (Abstract)
46. In der Beeck, M. Der Begriff der Vorgestalt in der Sprachentwicklung des Kleinkindes. *Nervenartz*, 1952, **23**, 464–466.
47. Johannes, T. Der Einfluss der Gestaltbindung auf das Behalten. *Arch. ges. Psychol.*, 1932, **85**, 411–457.
48. Johannes, T. Der Einfluss der Gestaltbindung auf das Behalten. 2 Teil. *Arch. ges. Psychol.*, 1939, **104**, 74–130.
49. Johnson, D. M. A modern account of problem solving. *Psychol. Bull.*, 1944, **41**, 201–229.
50. Johnson, D. M. *The psychology of thought and judgment.* New York: Harper, 1955.
51. Jung, C. G. *Studies in word association* (M. D. Eder, Trans.). New York: Moffat, 1919.
52. Jung, C. G. The psychology of dementia praecox. *Nerv. ment. Dis. Monogr.*, 1936, No. 3.
53. Kasanin, J. S. (Ed.) *Language and thought in schizophrenia: collected papers.* Berkeley: Univer. of California Press, 1944.
54. Klein, G. S. Perspectives to a research program on the organization of personality. Paper read at N. Y. Psychol. Ass., New York, January, 1954.
55. Klein, G. S., Spence, D. P., Holt, R. R., & Gourevitch, S. Preconscious influences upon conscious cognitive behavior. *Amer. Psychol.*, 1955, **10**, 380. (Abstract)
56. Kline, M. V., & Schneck, J. M. Hypnosis in relation to the Word Association Test. *J. gen. Psychol.*, 1951, **44**, 129–137.
57. Kohs, S. C. The association method in its relation to the complex and complex indicators. *Amer. J. Psychol.*, 1914, **25**, 544–594.
58. Kragh, U. *The actual-genetic model of perception-personality.* Lund: CWK Leerup, 1955.
59. Krueger, F. The essence of feeling: outline of a systematic theory. In M. L. Reymert (Ed.), *Feelings and emotions.* The Wittenberg Symposium. Worcester, Mass.: Clark Univer. Press, 1928. Pp. 58–78.
60. Kubie, L. S., & Margolin, S. The therapeutic role of drugs in the process of repression, dissociation, and synthesis. *Psychosom. Med.*, 1945, **20**, 147–151.

61. Lazarus, R. S. Subception: fact or artifact? A reply to Erikson. *Psychol. Rev.*, 1956, **63**, 343–347.

62. Lazarus, R. S., & McCleary, R. A. Autonomic discrimination without awareness: a study of subception. *Psychol. Rev.*, 1951, **58**, 113–122.

63. Lindemann, E., & Clarke, L. D. Modifications in ego structure and personality reactions under the influence of the effects of drugs. *Amer. J. Psychiat.*, 1952, **108**, 561–567.

64. Lindemann, E., & Malamud, W. Experimental analysis of the psychopathological effects of intoxicating drugs. *Amer. J. Psychiat.*, 1934, **13**, 853–881.

65. Mantell, U. Aktualgenetische Untersuchungen an Situationsdarstellung. *Neue Psychol. Stud.*, 1936, **13**, 1–96.

66. McCleary, R. A., & Lazarus, R. S. Autonomic discrimination without awareness. *J. Pers.*, 1949, **18**, 171–179.

67. McFarland, R. A. The psychological effects of oxygen deprivation (anoxemia) on human behavior. *Arch. Psychol.*, 1932, No. 145.

68. McGinnies, E. Emotionality and perceptual defense. *Psychol. Rev.*, 1949, **56**, 244–251.

69. Meadows, A. Anxiety, concrete thinking and blood pressure changes in schizophrenia. Unpublished doctor's dissertation, Harvard Univer., 1951.

70. Messer, A. Experimentell-psychologische Untersuchungen über das Denken. *Arch. ges. Psychol.*, 1906, **8**, 1–224.

71. Meumann, E. Über Assoziationsexperimente mit Beeinflussung der Reproduktionzeit. *Arch. ges. Psychol.*, 1907, 9, 116–150.

72. Mintz, A. Schizophrenic speech and sleepy speech. *J. abnorm. soc. Psychol.*, 1948, **43**, 548–549.

73. Mörschner, W. Beträge zur Aktualgenese des Gegenstanderlebens. *Arch. ges. Psychol.*, 1940, **107**, 125–149.

74. Ombredane, A. *L'aphasie et l'élaboration de la pensée explicite.* Paris: Presses Universitaires de France, 1951.

75. Osgood, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

76. Patrick, C. Creative thought in poets. *Arch. Psychol.*, 1935, No. 178.

77. Patrick, C. Creative thought in artists. *J. Psychol.*, 1937, **4**, 35–73.

78. Patrick, C. Scientific thought. *J. Psychol.*, 1938, **5**, 55–83.

79. Pena, C. A genetic evaluation of perceptual structurization in cerebral pathology: an investigation by means of the Rorschach test. *J. proj. Tech.*, 1953, **17**, 186–199.

80. Phillips, L., & Framo, J. L. Developmental theory applied to normal and psychopathological perception. *J. Pers.*, 1954, **22**, 465–474.

81. Pick, A., & Thiele, R. Aphasie. In A. Bethe (Ed.), *Handb. d. Norm. u. Pathol. Physiol.*, Vol. XV, 2. Berlin: Springer, 1931.

82. Rapaport, D. Consciousness: a psychopathological and psychodynamic view. In H. A. Abramson (Ed.), *Conference on problems of consciousness.* New York: Josiah Macy, Jr. Foundation, 1951. Pp. 18–57.

83. Rapaport, D. Toward a theory of thinking. In D. Rapaport (Ed.), *Organization and pathology of thought.* New York: Columbia Univer. Press, 1951. Pp. 689–730.

84. Rapaport, D., Gill, M., & Schafer, R. *Diagnostic psychological testing.* Vol. 2. Chicago: Year Book Publishers, 1946.

85. Razran, G. Salivation and thinking in different languages. *J. Psychol.*, 1935, **1**, 145–151.

86. Razran, G. Semantic, syntactic, and phonetographic generalization of verbal conditioning. *Psychol. Bull.*, 1939, **36**, 578. (Abstract)

87. Razran, G. A quantitative study of meaning by a conditioned salivary technique (salivary conditioning). *Science*, 1939, **90**, 89–90.

88. Razran, G. Semantic and phonetographic generalization of semantic conditioning to verbal stimuli. *J. exp. Psychol.*, 1949, **39**, 642–652.

89. Razran, G. Attitudinal determinants of conditioning and of generalization of conditioning. *J. exp. Psychol.*, 1949, **39**, 820–829.

90. Razran, G. Experimental semantics. *Trans. N. Y. Acad. Sci.*, 1952, **14**, 171–176.

91. Riess, B. F. Semantic conditioning involving the GSR. *J. exp. Psychol.*, 1940, **26**, 238–240.

92. Rinkel, M., DeShon, H. J., Hyde, R. W., & Solomon, H. C. Experimental schizophrenia-like symptoms. *Amer. J. Psychiat.*, 1952, **108**, 572–578.

93. Sander, F. Über räumliche Rhytmik. *Neue Psychol. Stud.*, 1926, **1**, 125–158.

94. Sander, F. Experimentelle Ergebnisse der Gestaltpsychologie. In E. Becher (Ed.), *10 Kongr. Ber. Exp. Psychol.* Jena: Fischer, 1928. Pp. 23–88.

95. Sander, F. Structures, totality of experience, and gestalt. In C. Murchison (Ed.), *Psychologies of 1930.* Worcester, Mass.: Clark Univer. Press, 1930. Pp. 188–204.

96. Sander, F., & Jinuma, R. Beiträge zur Psychologie des stereoskopischen Sehens. I. Mitteilung. Die Grenzen der binokularen Verschmelzung in ihrer Abhängigkeit von der Gestalthöhe der Doppelbilder. *Arch. ges. Psychol.*, 1928, **65**, 191–207.

97. Schafer, R. A study of thought processes in a word association test. *Charact. Pers.*, 1945, **13**, 212–227.

98. Schilder, P. *Mind: perception and thought in their constructive aspects.* New York: Columbia Univer. Press, 1942.

99. Schilder, P. On the development of thoughts. In D. Rapaport (Ed.), *Organization and pathology of thought.* New York: Columbia Univer. Press, 1951. Pp. 497–518.

100. Schilder, P. Studies concerning the psychology and symptomatology of general paresis. In D. Rapaport (Ed.), *Organization and pathology of thought.* New York: Columbia Univer. Press, 1951. Pp. 519–580.

101. Siipola, E., Walker, W. N., & Kolb, D. Task studies in word association, projective and nonprojective. *J. Pers.*, 1955, **23**, 441–459.

102. Silberer, H. Report on a method of eliciting and observing certain symbolic hallucination-phenomena. In D. Rapaport (Ed)., *Organization and pathology of thought.* New York: Columbia Univer. Press, 1951. Pp. 195–207.

103. Smith, F. An experimental investigation of perception. *Brit. J. Psychol.*, 1914, **6**, 321–362.

104. Smith, G. J. W., & Henriksson, M. The effect on an established percept of a perceptual process beyond awareness. *Acta Psychologica*, 1955, **11**, 346–355.

105. Smith, G. J. W., & Klein, G. S. Cognitive controls in serial behavior patterns. *J. Pers.*, 1953, **22**, 188–213.

106. Smith, W. W. *The measurement of emotion.* New York: Harcourt, Brace, 1922.

107. Sommer, W. Zerfall optischer Gestalten Erlebnissformen und Strukturzusammenhänge. *Neue Psychol. Stud.*, 1937, **10**, 1–66.

108. Speich, R. Reproduktion und psychische Aktivität. *Arch. ges. Psychol.*, 1927, **59**, 225–338.

109. Stein, M. I. Personality factors involved in the temporal development of Rorchach responses. *J. proj. Tech.*, 1949, **13**, 355–414.

110. Stockings, G. T. Clinical study of the mescaline psychosis with special reference to the mechanism of the genesis of schizophrenic and other psychotic states. *J. ment. Sci.*, 1940, **86**, 29–47.

111. Storch, A. The primitive archaic forms of inner experiences and thought in schizophrenia. *Nerv. ment. Dis. Monogr.*, 1924, No. 36.

112. Undeutsch, U. Die Aktualgenese in ihrer allgemeinpsychologischen und ihrer charakterologischen Bedeutung. *Scientia*, 1942, **72**, 37–42; 95–98.

113. Van Woerkom, W. Sur l'état psychique des aphasiques. *L'Encéphale*, 1923, **18**, 286–304.

114. Van Woerkom, W. Über Störungen in Denken bei Aphasiepatienten. *Mschr. Psychiat. Neurol.*, 1925, **59**, 256–322.

115. Varendonck, J. *The psychology of day dreams.* New York: Macmillan, 1921.

116. Viergutz, F. Das Beschreiben. Experimentelle Untersuchung des Beschreibens von Gegenständen. *Neue Psychol. Stud.*, 1933, **10**, 1–92.

117. Vigotsky, L. S. Thought in schizophrenia. *Arch. Neurol. Psychiat.*, 1934, **31**, 1063–1077.

118. Vinacke, W. E. The investigation of concept formation. *Psychol. Bull.*, 1951, **48**, 1–31.

119. Vinacke, W. E. *The psychology of thinking.* New York: McGraw-Hill, 1952.

120. Von Domarus, E. The specific laws of logic in schizophrenia. In J. S. Kasanin (Ed.), *Language and thought in schizophrenia: collected papers.* Berkeley: Univer. of Calif. Press, 1944. Pp. 104–114.

121. Voor, J. H. Subliminal perception and subception. *J. Psychol.*, 1956, **41**, 437–458.

122. Wallas, G. *The art of thought.* New York: Harcourt, Brace, 1926.

123. Wegrocki, H. J. Generalizing ability in schizophrenia: an inquiry into the

disorders of problem thinking in schizophrenia. *Arch. Psychol.*, 1940, No. 254.

124. Weinschenk, C. Der Begriff der Vorgestalt und die Hirnpathologie. *Nervenartz*, 1949, **20**, 355–361.

125. Weinschenk, C. Conrad's neuer Begriff der Vorgestalt und die Hirnpathologie. *Schweiz. Arch. Neurol. Psychiat.*, 1951, **67**, 101–118.

126. Wenzl, A. Empirische und theoretische Beiträge zur Erinnerungsarbeit bei erschwerter Wortfindung. *Arch. ges. Psychol.*, 1932, **85**, 181–218.

127. Werner, H. Studien über Strukturgesetze. IV. Über Mikromelodik und Mikroharmonik. *Z. Psychol.*, 1926, **98**, 74–89.

128. Werner, H. Studein über Strukturgesetze. V. Über die Ausprägung von Tongestalten. *Z. Psychol.*, 1927, **101**, 159–181.

129. Werner, H. Untersuchungen über Empfindung und Empfinden. II. Die Rolle der Sprachempfindung im Prozess der Gestaltung ausdruckmässig erlebter Wörter. *Z. Psychol.*, 1930, **117**, 230–254.

130. Werner, H. Musical "micro-scales" and "micromelodies." *J. Psychol.*, 1940, **10**, 149–156.

131. Werner, H. *Comparative psychology of mental development.* Chicago: Follett, 1948.

132. Werner, H. Microgenesis and aphasia. *J. abnorm. soc. Psychol.*, 1956, **52**, 347–353.

133. Werner, H., & Strauss, A. A. Pathology of figure-background relation in the child. *J. abnorm. soc. Psychol.*, 1941, **36**, 236–248.

134. Wever, E. G. Figure and ground in the visual perception of form. *Amer. J. Psychol.*, 1927, **38**, 194–226.

135. White, W. A. The language of schizophrenia. *Arch. Neurol. Psychiat.*, 1926, **16**, 395–413.

136. Wohlfahrt, E. Der Auffassungsvorgang an kleinen Gestalten. Ein Beitrag zur Psychologie des Vorgestalterlebnisses. *Neue Psychol. Stud.*, 1932, **4**, 347–414.

137. Woodworth, R. S. *Experimental psychology.* New York: Holt, 1938.

138. Woodworth, R. S., & Schlosberg, H. *Experimental psychology*, (Rev. Ed.). New York: Holt, 1954.

139. Wylie, R. C. Generalization of semantic conditioning of the galvanic skin response. Unpublished master's thesis, Univer. of Pittsburgh, 1940.

Received August 13, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 3, 1957

CHARACTERISTICS OF TYPE CONCEPTS WITH SPECIAL REFERENCE TO SHELDON'S TYPOLOGY

LLOYD G. HUMPHREYS[1]
Air Force Personnel and Training Research Center, Lackland Air Force Base, San Antonio, Texas

The writer became interested in Sheldon's physical and temperamental types (**10, 11**) because they have been so widely, and frequently so favorably, discussed in recent years. Relatively little investigation was needed in order to discover that the favorable discussions had little foundation in fact for the attitude expressed and that the use of Sheldon's types in further research should be discouraged.

In the course of this investigation interest was aroused concerning type theory generally. The conclusions reached have implications beyond the Sheldon types. It is believed, in brief, that traditional type[2] theories have important characteristics in common that arise inevitably from the definition of type. It is further believed that these characteristics make type concepts unsuited for most research purposes.

Organization of Discussion. A brief review of Sheldon's work will first be presented, including the logical-statistical analysis of his data that led the writer to reject his concepts. This will be followed by a discussion of the characteristics of type concepts and the similarities of types to ipsative (or relative) scales. Types, and ipsative scales, will then be contrasted with traits, or normative scales, and the applications of each in research pointed out. Finally the application of the multiple discriminant function to problems that in the past led to efforts at typing will be briefly described.

REVIEW AND ANALYSIS OF SHELDON'S CONCEPTS

Sheldon has described three physical types: endomorphs, characterized by visceral development; mesomorphs, characterized by skeletal and muscular development; and ectomorphs, characterized by neural development. Each of these three types can be reliably rated on a seven-step scale for every individual. Precise physical measurements can be used as the basis for these ratings. He has also described three temperament types; viscerotonic, somatotonic, and cerebrotonic. Each individual can also be reliably rated on a seven-step scale for each of the three temperament types. These ratings are obtained from a 60-item rating scale divided into three clusters of 20 traits each. An individual's type scores are conventionally written as three numbers, each having a theoretical range from one through seven, separated by

[1] Paper completed while serving as visiting professor, University of Illinois, fall semester, 1955, and while on leave from Personnel Laboratory, Air Force Personnel and Training Research Center, Lackland Air Force Base, San Antonio, Texas. The opinions or conclusions expressed herein are those of the author. These are not to be construed as necessarily reflecting the endorsement of the Department of the Air Force or of the Air Research and Development Command.

[2] The reader should beware reading into this discussion his own connotations with types. As the argument develops, it will be seen that a rather special definition of type emerges. This definition best characterizes Sheldon's types, but it is believed that it is generally applicable to his predecessors, such as Kretschmer, as well.

hyphens; e.g., 7-1-1, 4-4-3, 2-2-6, etc. Data are also presented apparently showing that physique and temperament are opposite sides of the same coin; i.e., the correlations between the logically related types of physique and temperament are all in the neighborhood of .80.

Usual assessment of Sheldon's contribution. The assessment of Sheldon's contribution to typology is frequently divided into two parts. In the first place, he is credited with the introduction of quantification in typing procedures. The attitude taken in this paper, however, is that there is no virtue in quantification if there is no justification for the variables measured. Thus, thorough investigation of the origin and characteristics of his variables is indicated. In the second place, Sheldon is credited with obtaining the most substantial relationships yet obtained between physique and temperament. In evaluating this second contribution, it is more important to evaluate the controls used in obtaining the measures correlated than it is to compute the standard errors of those correlations.

Analysis of Sheldon's types of physique. The error involved in accepting Sheldon's work at face value becomes apparent when his procedure is reviewed. First, with regard to establishing the physical types, it is clear that the procedure was not empirically sound. The types originated in the arm chair. Sheldon did have large numbers of photographs spread out before him when he selected the types, but that hardly makes the procedure empirical. If Thurstone had spread 56 printed tests before himself and decided what factors were needed to describe performance on these tests, he would have produced a set of human abilities with about as much justification as Sheldon has for his physical types. In a situation of this sort, it is unlikely that the observer would find much beyond what he expected to find. Any similarity between the Sheldon and Kretschmer types is certainly not coincidental, and does not necessarily mean that Kretschmer was groping in the right direction in a primitive sort of way.

An analysis of the intercorrelations of Sheldon's types will also contribute to the evaluation of his concepts. Sheldon (**10**) has published the intercorrelations of his physical types, based on two samples of 2,000 and 200 cases respectively. He also included, in an appendix, data for 4,000 cases. In order to obtain greater sampling stability, correlations were computed for the sample of 4,000 cases by the writer. These are presented in Table 1. In comparing them with those published by Sheldon (the sample of 2,000 cases was presumably included in the 4,000), an additional advantage of the new computations is discovered: what appears to be a computational error in the published value for the correlation between endomorphy and ectomorphy in the sample of 2,000 cases is corrected. A value of −.27 is improbably low in comparison to the present value of −.40. It might also be noted here that other errors have been found in Sheldon's computations (**7**).

TABLE 1

INTERRELATIONSHIPS OF PHYSICAL TYPES

	Endomorphy	Mesomorphy	Ectomorphy
Endomorphy	.765	−.300	−.402
Mesomorphy		.818	−.576
Ectomorphy			.833

Note.—Intercorrelations of types are presented in the usual fashion. Multiple correlations between each type and the other two are listed in the diagonal. $N = 4{,}000$.

Table 1 also includes, as the diagonal entries, the multiple correlations between each possible pair of types and the third. It is seen that the multiples are in every case much higher than the zero-order correlations, but cannot be said to approximate unity. Sheldon makes a good deal of this "thickness," i.e., evidence for three-dimensionality. Before accepting this evidence for three-dimensionality, however, other factors must be considered. Ekman (**4**) has also considered these factors in evaluating the claim for three dimensions. The present development differs from his, particularly with respect to the use of the multiple correlation technique.

TABLE 2

Comparison of Sheldon's Scales with the Converted Scales

Endomorphy			Mesomorphy			Ectomorphy		
Sheldon's Scale		Converted Scale	Sheldon's Scale		Converted Scale	Sheldon's Scale		Converted Scale
7	—	10	7	—	8.5	7	—	8
6	—	8	6	—	7	6	—	6.5
5	—	6	5	—	5.5	5	—	5
4	—	4	4	—	4	4	—	4
3	—	3	3	—	3	3	—	3
2	—	2	2	—	2	2	—	2
1	—	1	1	—	1	1	—	1

Note.—The new scales, determined by inspection, were designed to make the regressions of each type on the others more nearly linear.

The scatter plots presented by Sheldon not only represent negative correlations, but show evidence of a good deal of curvilinearity as well. It is possible to correct for some of this distortion. The transformations shown in Table 2 were obtained by estimating the amount the scales needed to be "stretched" at the high end in order to convert the curvilinear regressions to something approaching linearity. The rationale for a correction of this type, as Sheldon has stated, may be the lack of linear relationship between stimulus units and judgments of equal increments. The intercorrelations were then recomputed, and new multiple correlations determined. These are presented in Table 3. As compared to Table 1, a gratifying increase in the multiples is evident, but one is still uncertain whether any one variable is completely determined by the other two.

There are other legitimate corrections that can be applied as long as we are interested in the problem of intrinsic relationships among the types. Certain errors of measurement, which are involved in the determination of physical type, and errors of grouping, since the scales are not in actuality continuous, also attenuate the relationships obtained. By applying Shepherd's correction to the standard deviations before computation of *r*, correction was made for the second of these attenuating factors. These relationships are presented in Table 3. Assuming reliability coefficients of both .95 and .97,

TABLE 3

Type Interrelationships After Correction for Curvilinearity and Discontinuity

	Endomorphy	Mesomorphy	Ectomorphy
Endomorphy	.835/.901	−.347	−.444
Mesomorphy	−.334	.870/.924	−.613
Ectomorphy	−.424	−.586	.881/.931

Note.—Correlations involving the converted scales are below the diagonal. Correlations above the diagonal were computed after applying Shepherd's correction to the standard deviations. Multiples are in the diagonal with the higher of the two representing the relationships obtained after making both corrections. $N = 4,000$.

TABLE 4

Type Interrelationships After Correction for Unreliability

	Endomorphy	Mesomorphy	Ectomorphy
Endomorphy	.953/.989	−.365	−.467
Mesomorphy	−.358	.964/.992	−.645
Ectomorphy	−.458	−.632	.968/.993

Note.—The data, after correction for attenuation, are presented as in Table 1. Reliabilities were assumed to be first, at the .95 level and, second, at .97. Values based on the former assumption are above the diagonal; others, below. The separate sets of multiple correlations obtained are again in the diagonals and are approximately unity. $N = 4{,}000$.

the values of r in Table 3 were corrected for errors of measurement, and new multiples were computed. These results are presented in Table 4. With assumed reliabilities of .95, the evidence for three-dimensionality completely disappears. The higher values leave little room for a third dimension. We can conclude that Sheldon has evidence for no more than two independent (not necessarily valid) types of human physique. Ekman's conclusion is thus thoroughly substantiated.

Origin of the temperament types. Sheldon's procedure in establishing the temperament types is also subject to criticism. Sheldon states that in selecting the 20 traits used to describe each of the three types he used a procedure similar to factor analysis. An impartial critic would prefer the term "cluster analysis," and one would add "statistically naive" as well. Sheldon's statistical criteria for trait selection were as follows: correlations of at least +.60 between all traits within each cluster, and correlations of at least −.30 with all traits in the other two clusters. He states that on a priori grounds he expected to find four clusters and was surprised to find only three. The a priori reasons were evidently not statistical in nature. As long as adequate numbers of cases were used (to avoid gross sampling errors), his criteria for selection made it statistically impossible to obtain more than three clusters of traits. Furthermore, it became equally certain that any two would approximately determine the third.

The statistical argument here is simple. Since each trait in a cluster had to be correlated to the extent of at least −.30 with every trait in the other clusters, the mean correlations between single traits in separate clusters must be substantially greater than −.30. The correlations between clusters will be still greater since sums of 20 traits are correlated. It is estimated, on the basis of the formula for the correlation of sums, that the correlations between clusters would approach −.50. It is statistically impossible to find more than three variables with intercorrelations in this neighborhood, since at this point the multiple correlation between any two and the third is unity. For four variables, the intercorrelations would have to be as low as −.333, an obvious impossibility starting with the a priori criteria of trait selection used by Sheldon.

Correlations published by Sheldon (11) for a sample of 200 cases for the temperament types are presented in Table 5. Multiples, compiled by the writer, appear in the diagonal as before. These are sufficiently high that it seems useless to go through the series of corrections made on the data for physical types. The less reliable nature of the temperament ratings is

TABLE 5

Interrelationships of Temperament Types

	Viscerotonia	Somatotonia	Cerebrotonia
Viscerotonia	.815	−.34	−.37
Somatotonia		.873	−.62
Cerebrotonia			.875

Note.—Intercorrelations of the types, taken from Sheldon, are presented in the usual fashion. Multiple correlations between each type and the other two are listed in the diagonal. No corrections have been applied to the data. $N = 200$.

in itself probably sufficient to account for the obtained values being less than unity. We can safely conclude that Sheldon has evidence for no more than two independent (not necessarily valid) types of temperament.

Physique-temperament correlations. The published correlations relating the physique and temperament types, while undoubtedly higher and more stable from the sampling point of view than any others in the literature, are basically defective. Several reviewers who should have known better have disregarded the fact that the same person (Sheldon) made the ratings of both temperament and physique. Sheldon at least recognized the danger in this procedure, but discounted it for two reasons: he states that he recognized the difficulty at the time the ratings were being made, and the ratings of temperament preceded the ratings of physique. These arguments are not convincing. The relationships in question could be completely invalidated by this aspect of the procedure. The only legitimate conclusion to be drawn concerning these relationships is "not proven." Let the reader reflect for a moment how the data from an analogous situation would be received by a group of biologically oriented psychologists. A social psychologist has an hypothesis concerning the effects of democratic and autocratic home atmosphere on the development of a certain personality trait. His research involves a correlation between ratings of homes and ratings of the personality trait, both made by himself. A high correlation is found. The analogy seems sufficiently close, and the conclusion so apparent, as to need no further comment.

Sheldon's more recent report on juvenile delinquency (**12**) also shows evidence of inadequate control. The conclusion that physical type is highly related to juvenile delinquency is based on a comparison of his delinquent sample with his college undergraduates.

Evaluation of Sheldon's typology. Sheldon's claims for having established relationships between physique and temperament are thus "thrown out of court" for lack of evidence. More basic, however, is the doubt cast on the validity of his type concepts. His temperament types were arbitrarily determined by the statistical criteria. His physical types arose from the arm chair and were undoubtedly influenced by the same line of statistical reasoning. Research workers, if they wish to make use of Sheldon's types, are advised to discard one physical type and the corresponding temperament type. This would result in savings of measurement time and statistical analysis of data. If multiple regression analysis is planned, however, the recommended procedure becomes compulpulsory. Beta weights can be reliably determined on only two of three mutually dependent variables.

Even if the research worker in this field discards one of the three types, he can still have no confidence in the meaningfulness of the two retained. This is not to say that empirical relationships with the types cannot be obtained, though Eysenck's review (**5**) indicates that few have been established. The more careful investigation of factors in delinquency by Glueck and Glueck (**6**) does indicate some nonchance relationship with body build. The possibility remains, and will be discussed in greater detail later, that a more sophisticated approach to the problem would produce more substantial relationships in those cases where some relationship has been shown.

The Logic of Type Variables

With the completion of these statistical analyses of Sheldon's types, the writer became interested in the logic of type variables as they have been used historically by Kretschmer and others. This logic apparently explains some of Sheldon's mistakes, i.e., his interest in negative correlations. It also furnishes reasons why a priori types should be discarded. The argument here will again be found to parallel in part a theoretical development of Ekman (**3**). The latter did not see, however, that his reasoning applied as well to Sheldon as to Kretschmer.

Definition of type. A type has traditionally been defined in terms of an ideal person. A type score is the degree to which a given individual approaches the ideal. Ideals (types) are defined in terms not only of the presence of certain traits to high degree, but also of the virtual absence of all other traits. The description of a second ideal (type) will involve high scores on certain traits and low scores on others that entered the description of the first ideal (type) in opposite degree. Negative correlations among types naturally follow, and the smaller the number of types deemed necessary to encompass the range of human differences, the higher will be these negative correlations.

The scatter plots of the correlations among Sheldon's types are of interest in this regard beyond the evidence for curvilinearity of regression discussed earlier. These plots take the form of a "T" tilted to the right; i.e., there are no entries above the upper left or lower right diagonals. Two high type scores, e.g., 7-7-1, 7-1-7, or 1-7-7, are impossible because two ideals cannot both be approximated in one individual. A maximum score for one type also assures two minimum scores for the other types, e.g., 7-1-1, 1-7-1, or 1-1-7, since the one high score means that the individual is low on all the other traits that in various combinations determine the remaining types.

The definition of type thus far evolved is, however, lacking in one particular. It does not explain why three low type scores are not found, nor why one average score is always accompanied by at least one other average score. The definition does not account, in other words, for the fact that the three type scores add up to a constant, as has been shown earlier.

It seems to have been assumed by Sheldon that the definition of the ideal (type) should be in relative terms. Approximation to the ideal mesomorph, for example, does not depend on absolute height or weight, but is a function of relative bodily proportions. Defined in this way the physique of every individual is completely described by the types selected. By definition there are no 1-1-1 individuals. There is nothing remarkable in the fact that certain combinations of scores do not occur in nature as Sheldon implies. Certain combinations are prohibited by the nature of the concepts selected to describe human physique or temperament.

Examples of types. It may be useful at this point to give several examples of the operation of type concepts. These illustrations will serve to clarify further the definition of type. They will also serve as the main argument concerning the arbitrariness of number of types used by any one theorist.

Let us suppose that there are twelve observable human mental abilities. It would be possible for a type theorist in viewing this particular range of human differences to establish arbitrarily only two ability types. One could, for example, speak of the intellectual and mechanical types. The first would be defined by the presence of high scores on about half of the 12 abilities, low scores on all of the rest. High and low are of course defined relative to the person's own mean. Any exception to this pattern would reduce the size of the type score, i.e., the perfect intellectual type is low in any trait required for mechanical occupations. The second type would be defined by the opposite combination of abilities; i.e., low on the group where intellectuals are high, high on all of the rest. The resulting correlation between the two types would be −1.00. Note that every individual can now be placed at some point along each of these type scales, but that knowledge of one determines the other, and that as long as low and high are rated relative to the individual's own profile of abilities, the sum of the two type scores will be a constant for all individuals.

Three ability types could equally well be established. One type could be defined in terms of high scores on about a third of the abilities, low on the rest. The other two types would be defined in a similar manner. In this case the correlations between the three types will be of the magnitude of −.50. Again, all individuals can be placed at some point along each of these scales, with any two defining the third. Three types now encompass the entire range of human abilities.

This process could obviously be continued until 12 types were defined. We would still find negative correlations between types, averaging about −.09 for this number of types. Twelve types would also make possible a larger number of combinations of type scores, including several fairly high scores for any one individual. We would still find, however, that the top possible score for one type would force the other scores to minimum levels. In general, the same biasing factors would be present, but their force would be somewhat dissipated by the larger number of degrees of freedom available. Note that, no matter what the correlations might be between the trait measures, from past experience we know that there would be a range of positive values—the correlations among the type scores would necessarily be negative because of the way in which types are defined.

It will be remembered that Sheldon strove for negative correlations among his types of temperament and that he was pleased with negative correlations among his types of physique. It is now seen that he was following the logic of the traditional type concept. Seeking high negative correlations automatically produces a small number of types. Using relative standards ensures that everyone will have a high score some place. These characteristics, combined with a presumed high degree of generality in explaining human behavior (as a matter of fact the presence of a pigeonhole for everyone is frequently assumed to constitute evidence for the generality), make type concepts well nigh irresistible for the clinically oriented person. Although Sheldon's predecessors did not quantify their types, their theories had basically these same characteristics.

Substitutes for Types

At the point where there are as many types as there are measured traits, types become simply ipsative (**1**) or relative (**2**) scales. Score values are obtained with reference to the person's own mean. Another common way of stating the same thing is that level has been removed from the profile. Sheldon's types could also be characterized as ipsative scales, although they differ from the usual scales of this sort in their complexity, i.e., a single type is defined by many facets of the person. Otherwise the parallel is complete. Intercorrelations among any number of ipsative scales will tend to be negative as long as these scales are obtained from a common score matrix. The size of the negative correlations will be a function of the number of ipsative scales. No one can be good on everything on such scales. Everyone will be high in something, low in something else. All persons' scores will add to a constant which will be equal to the amount, added to a standard score of zero, found necessary to avoid negative scores on any one scale.

Traits, in contrast to types, have historically been associated with normative scales. The contrast between the functional characteristics of traits and types, or normative and ipsative scales, is marked. A proposed trait measure may have correlations with other normative scales ranging from plus to minus unity. The average intercorrelation of a group of trait measures may be anything in the same range. As long as correlations with other measures are not unity, all possible combinations of trait scores will appear. Relative frequencies of such combinations will vary, of course, but only as a function of the correlations between the scales and the shapes of the marginal distributions. The contribution to variance of across-the-board differences between individuals can be large or small relative to the differences within individuals. No matter how low the intercorrelations of trait measures are, however, there will be some few persons low on everything, others high on everything. The statement found in many elementary texts, "correlation not compensation," holds for traits; but the opposite statement, "compensation not correlation," holds for types.

Another difference between traits and types is that the trend in trait measurement has been toward more specificity. Thus the complex trait measure of general intelligence has been giving way to factor measures of separate aptitudes. If type-like scales are desired for research purposes, it would be useful to give up complex types such as those of Sheldon and use specific ipsative scales in sufficient number to cover the area of interest.

Choice of scale. For prediction purposes an investigator must choose the type of scale which is fitted for the problem at hand. In most cases this will be a normative scale. Most proficiency criteria, for example, are themselves normative. It is highly doubtful that Sheldon's types, or other more carefully selected ipsative scales, can predict athletic achievement as well as normative scales of physique. One might be able to define the line-backer type, but if a given example weighed 120 pounds he probably would not be suitable material for the college team. In the same way, normative aptitude scales are better predictors of academic achievement than ipsative scales.

The writer is not at all certain that there are occasions when an ipsative scale would be preferred. This possibility should not be ruled out, however, without thorough exploration. A good bet for the tryout of ipsative scales is in the prediction of decisions. Such criteria would seem to result from the balancing of tendencies (traits) within the individual, not from his standing in a group on the several traits.[3]

It was mentioned earlier that Glueck and Glueck had found a nonchance relationship between somatotype and delinquency. It is possible that this criterion is also basically ipsative. But whatever the nature of the criterion, it is highly probable that greater differentiation could have been accomplished by an empirical combining of several specific measures. The discriminant function could be applied to either ipsative or normative scales, or a combination of the two. The choice between scales would be made empirically in terms of the differentiation obtained.

The multiple discriminant function. As a matter of fact, the discriminant function, or better, the multiple discriminant function (**13**, **14**), is a logical technique to substitute for typing procedures. Is person A most like the average artist, salesman, physician, engineer, or lawyer? Is person B most like runners, jumpers, or shot-putters? Is person C most like a brain injury case, a schizophrenic case, or a psychopathic deviate? Although a given discriminant may have characteristics reminiscent of types, there is a basic difference—it is formed to answer a specific problem. One would rarely if ever wish to use a discriminant successful in one area of research for another problem. Trying out complex types in each new problem area is equally unjustified and is equally to be discouraged.

Steps in scale development. It is clear to the writer that methodological research on scale development should proceed in accordance with a logical order, which applies to physique as well as to interest, temperament, and aptitude.

1. The first priority should be given to the development of adequate normative scales. These scales should be fairly specific and relatively homogeneous, though the scalability criterion of homogeneity should not be generally applied. Requiring too high a degree of homogeneity results in too many scales giving too little useful differential information. The direct method of factor analysis may be a useful tool in the development of the required normative scales.

2. Ipsative scales should usually follow the normative. For one thing this would allow a rational choice of a normative group of scales for development of the ipsative scoring. Mainly, however, one must know the characteristics of the normative scale before the ipsative scale can be given meaning. A possible exception to this order of development is in the use of the paired comparison item format,

[3] It should be noted that tryout of ipsative scales has value only for theoretical purposes. William V. Clemens has shown in an unpublished report from the University of Washington (October 1956) that the group of normative scales from which the ipsative scales are developed will always give a multiple correlation with an outside variable as high or higher than the latter. Thus from the point of view of prediction, ipsative scales are unnecessary. Appropriate combinations of positive and negative weights of the normative scales can always accomplish the same result.

or one of its derivatives, for interest and personality measurement. An initial normative scale may not be essential for an adequate ipsative scale in these areas, though we should not rest content to use paired comparison scales in place of adequate normative scales in situations requiring the latter.

3. With separate scales well in hand, it is appropriate to consider combinations for the prediction of assorted criteria. It also seems clear that this should be done both empirically and statistically; i.e., the combination should be for a particular purpose, it should follow—not precede—measurement, and it should be computed by an appropriate formula. These criteria rule out use of Sheldon's types. The last rules out clinical methods of combining test information.

4. Consideration of statistical methods of combining should not be limited to the multiple regression procedure. The applicability of the discriminant function to problems associated in the past with types has already been described. Other prediction problems may yield to pattern analysis techniques (**8, 9**). Note that these techniques are not typing procedure as used in this discussion. Pattern analysis is here considered as another way of combining data to solve a particular problem.

Summary and Conclusions

Sheldon's physical and temperamental types, and their joint relationship, have been critically examined. A number of limitations of his research are apparent. The physical types originated in the arm chair. Measurement entered later as a means of differentiating objectively the subjectively determined types. It was also shown that the choice of types to describe human physique and temperament automatically restricts the data in predictable ways. Of necessity, type interrelationships are negative, and certain combinations of scores are prohibited by the nature of the concepts. As a result of the expected mutual dependence of types, there is evidence in Sheldon's data for no more than two independent types, either of physique or temperament. The research worker, if he uses these types, is therefore advised to discard one type of physique and its temperament counterpart in his investigations. Finally, the correlations relating physique to temperament are invalidated by the fact that the same judge (Sheldon) was responsible for both sets of ratings.

With respect to type concepts generally, it was suggested that types have traditionally been defined as mutually exclusive ideals. Thus, two types can never be represented in high degree in one person. Furthermore, types have been defined by relative measures so that no one is low in everything; i.e., a pigeonhole is provided for everyone. This tends to give type concepts a spurious degree of attractiveness. The size of the complex involved in the type is arbitrary, however, so that the number of types can vary from two up to the number of discriminable traits. Each such set of types has the same characteristics, but the average level of negative correlations decreases as the number of types increases. An increasing number of types also allows more degrees of freedom for various combinations of type scores, but certain combinations will still be prohibited in even large numbers of types.

When the number of types is equal to the number of traits, a type becomes an ipsative scale. Traits, on the other hand, are normatively scaled. Normative scales, in contrast to ipsative scales, do not bias intercorrelations and allow all possible score combinations. Traits are recommended for most predictive purposes, since most criteria are themselves normative. Specific ipsative scales may be useful for certain problems, though this question is largely unexplored.

In place of a priori complex types, the use of the multiple discriminant function is recommended for problems traditionally associated with typing. Discriminants may have properties similar to types, but always differ in one important particular—they are computed to solve a particular problem. Still other problems traditionally associated with type concepts may yield to pattern analysis techniques.

For those investigators interested in problems of physique, it is recommended that they start with the trait approach and within reason exhaust its possibilities. One might wish subsequently to explore ipsative scoring of the separate traits. Finally, for specific prediction problems, various possible mathematical combinations of either normative or ipsative scales should be tried.

REFERENCES

1. Cattell, R. B. *Description and measurement of personality.* Yonkers-on-Hudson: World Book, 1946.
2. Coombs, C. H. *A theory of psychological scaling.* Ann Arbor: Engineering Research Institute, Univer. of Michigan, 1951.
3. Ekman, G. On typological and dimensional systems of reference in describing personality. *Acta psychol.*, 1951, **8**, 1–24.
4. Ekman, G. On the number and definition of dimensions in Kretschmer's and Sheldon's constitutional systems. In *Essays in psychology dedicated to Daniel Katz.* Uppsala: Almquist, 1951.
5. Eysenck, H. J. *The structure of human personality.* London: Methuen and Co., 1953.
6. Glueck, S., & Glueck, Eleanor. *Unravelling juvenile delinquency.* New York: Commonwealth Fund, 1950.
7. Lubin, A. A note on Sheldon's table of correlations between temperamental traits. *Brit. J. Psychol., Stat. Sect.*, 1950, **3**, 186–189.
8. Lubin, A., & Osburn, H. G. Some theorems on the use of pattern scoring for predictors of quantitative criteria. *Amer. Psychologist*, 1955, **10**, 413. (Abstract)
9. McQuitty, L. L. *A method of pattern analysis for isolating typological and dimensional constructs.* San Antonio, Texas: Headquarters Air Force Personnel and Training Research Center, Lackland Air Force Base (Research Report AFPTRC-TN-55-62), 1955.
10. Sheldon, W. H. *The varieties of human physique.* New York: Harper, 1940.
11. Sheldon, W. H. *The varieties of temperament.* New York and London: Harper, 1942.
12. Sheldon, W. H. *Varieties of delinquent youth.* New York: Harper, 1949.
13. Tiedeman, D. V., Bryan, J. G., & Rulon, P. J. *Application of the multiple discriminant function to data from the airman classification battery.* San Antonio, Texas: Headquarters Human Resources Research Center, Lackland Air Force Base (Research Bulletin 52-37), 1952.
14. Tiedeman, D. V., Rulon, P. J., & Bryan, J. G. The multiple discriminant function—a symposium. *Harv. educ. Rev.*, 1951, **21**, 71–95.

Received July 31, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 3, 1957

RELIABILITY AND BEHAVIOR DOMAIN VALIDITY: REFORMULATION AND HISTORICAL CRITIQUE

ROBERT C. TRYON
University of California

If an investigator should invent a new psychological test and then turn to any recent scholarly work for guidance on how to determine its reliability (e.g., **6**), he would confront such an array of different formulations that he would be unsure about how to proceed. After fifty years of psychological testing, the problem of discovering the degree to which an objective measure of behavior reliably differentiates individuals is still confused.

This confusion stems from a rigid adherence to unobjective and unrealistic postulates about the nature of measurement—assumptions originally invented by Spearman and William Brown a half century ago. We will review the unsuccessful efforts of psychologists over fifty years to free themselves from these restrictive orthodoxies. On the positive side, we will conceptualize and reformulate the problem in terms of the realities of objective measurement. Once an analyst has assessed the structure of a test, he can in most cases calculate the value of its reliability coefficient from the statistical constants of the test-samples (items) that compose it. The welter of different "methods" of calculating the reliability coefficient commonly employed are either different computational forms that yield the same correct value, or they refer to various empirical designs devised to estimate this value. A parallel confusion exists over determining the communality and cluster domain validity, treated elsewhere by the writer (**26**).

THE OBJECTIVE OPERATIONS OF MEASURING INDIVIDUAL DIFFERENCES IN ANY BEHAVIOR

The behavior analyst's first step in constructing a "test" is to conceptualize some property, X, of a group of individuals. If the individuals are men, X may be an ability like vocabulary knowledge, or some personality characteristic like rigidity. If they are rats, the property X may be maze learning. If Drosophila, X may be the geotropic reaction.

The second step is to define the property X in terms of *objective* specifications that directly lead to the taking of *test-sample* observations, $X_1, X_2, \cdots, X_n$, believed to elicit the defined behavior, X. These test-samples may be vocabulary items, ratings, entrances into blind alleys, vertical movements in a test tube.

The third step is to compute for each individual its *composite total score*, X_t, which is the sum of $X_1, X_2, \cdots, X_n$, i.e.,

$$X_t = X_1 + X_2 + \cdots + X_n. \qquad [1]$$

The analyst needs to know the *within individual variance*, or "error of measurement," of the total score, X_t. This individual variance would be the variability of the individual in many scores comparable to the observed composite, X_t. Any one of such comparable scores, say X_t', would also be composed of the sum of n test-samples, thus,

$$X_t' = X_1' + X_2' + \cdots + X_n', \qquad [2]$$

in which the primed test-samples are conceptualized as being of the same kind as those in the observed composite, X_t. The correlation between the observed X_t and the comparable construct, X_t' is called the ***reliability coefficient***, r_{tt}, of the observed composite. From the reliability coefficient the individual variance can, as we shall see later, be estimated.

Current practice in computing r_{tt} is variable. Some analysts insist that X_t' be an actual "comparable form" to X_t. Others prefer the "split-half" method, with Spearman-Brown correction for double length. Still others may compute r_{tt} from the variances of the observed test-samples (e.g., by the Kuder-Richardson formula). Some may even take r_{tt} to be the "test-retest" correlation.

Prevailing Assumptions About Measures of Individual Differences

Virtually all writers assert that the use of these important formulations follow from certain assumptions about the n test-samples that make up the observed composite, X_t. We shall examine these assumptions below, whence it will be obvious that rarely can it be shown objectively that test-samples do satisfy the requirements. This paper shows that such assumptions are not needed. In the last section of the paper we will see that despite the irrelevance and immateriality of these assumptions they have continued to govern most of the thinking about this problem since the inception many years ago of the Spearman-Yule theory of true and error factors on the one hand, and of the Brown-Kelley theory of statistically equivalent test-samples on the other.

The Spearman-Yule theory of true and error factors. We shall merely state this postulate here, later in the paper surveying its history. Spearman presented it in principle in 1910 (**18**), Yule participating to the extent of communicating a precise formulation of it to Spearman in 1908 (**27**, Ch. 11, ref. 7). And 44 years later Guilford accepts it as the "rationale" of psychological testing in his 1954 *Psychometric Methods* (**6**).

In brief, the theory asserts that for scores on any two raw test-samples, X_i and X_j, each deviation score, x_i and x_j, of an individual, is determined by an "underlying" true factor, x_∞, plus an "error" factor, e. The error factors of x_i and x_j are postulated as being uncorrelated with x_∞, and with each other. In short,

$$\left.\begin{aligned} x_i &= x_\infty + e_i \\ x_j &= x_\infty + e_j \end{aligned}\right\} \qquad [3]$$

$$r_{e_i x_\infty} = r_{e_j x_\infty} = r_{e_i e_j} = 0. \qquad [4]$$

Spearman thought of x_∞ as "g," a general factor running through all cognitive abilities, but in the eyes of modern factor analysts "g" is usually replaced by a composite of more than one common factor plus a factor specific to each test-sample. The true factor, x_∞, is released from its "underlying" status by some writers who conceive it to be the mean of many test-sample scores, but even in this conception the errors are postulated as being uncorrelated. The errors are furthermore postulated as operating equally in the test-samples, the net effect being that the test-samples reveal equal variances and equal inter-r's. All the standard formulas for computing the reliability of a test can be derived from these postulates (**7**, Ch. 2; **15**). So derived, however, these formulas would be restricted in use only to those measures for which it could be demonstrated that these postulates have substantive support—an unattractive requirement for formulas so generally employed.

The Brown-Kelley theory of statisti-

cally equivalent test-samples. The other conception, presented by William Brown (**1**, **2**) also in 1910, later systematically formulated by Truman Kelley (**13**), is the basic doctrine generally accepted 40 years later by Gulliksen in his 1950 *Theory of Mental Tests* (**7**, Ch. 3 ff.). It postulates that all the n test-samples in X_t must have equal standard deviations, and equal intercorrelations, i.e.,

$$\sigma_i = \sigma_j = \sigma, \text{ a constant} \quad [5]$$

$$r_{ij} = r, \text{ a constant.} \quad [6]$$

In short, this conception ignores "underlying factors" but accepts the equivalence of test-samples. All the standard formulas for calculating the reliability and domain validity of a composite test can also be derived on these assumptions (**7**, Ch. 3; **13**).

Were we to restrict the use of these formulas only to tests whose test-samples met these conditions of strict equality of σs and inter-rs the formulas would obviously not be applicable to most of the commonest situations, such as those in which the test-samples are true-false items with differing proportions of true responses.

Objective Principles of Domain Sampling

No restrictive postulates or assumptions about the *observed* test-samples are in fact required in the development and use of the standard formulations of reliable individual measurement. The standard formulas follow directly from the operations employed in objectively sampling behavior. Postulates of "underlying factors" are superfluous, and test-samples may have different variances and covariances.

In computing the reliability coefficient of the total score, X_t, the analyst is seeking an answer to this question: What is the value of the correlation, r_{tt}, between the observed X_t scores and a second set of composite scores, X_t', earned on a "comparable form" of the X_t composite?

Comparable form, X_t', as a construct. We must define comparability of the X_t' composite in terms of the realities of the observed X_t composite, as follows: *A comparable X_t' composite is one whose n test-samples vary on the average as much in σs and inter-rs as do the n test-samples in the observed X_t composite.* Analysts may not anticipate actually setting up such a comparable second composite in order to calculate r_{tt}. Indeed, it may not be feasible to do so. Furthermore, with the exception of certain "stratified composites" to be discussed later, it is unnecessary to do so, for the second X_t' composite is a construct whose average statistical properties are *by definition* those of the observed X_t composite at hand. This construct X_t' composite is in fact a criterion by which to determine the degree to which an actual second composite is comparable. If such an actual second composite reveals the same average properties as the first, then it *is* comparable to the first by definition; if its properties deviate from the first, then it is *not* comparable.

To make the matter concrete, look at the data in Table 1. In the score matrix at the top left you see five observed test-sample scores of 10 actual individuals. Individual 1, for example, has the five X_i scores, 6, 2, 1, 0, 0 which add up to a composite X_t score of 9. The problem is to calculate the reliability coefficient of the set of X_t scores of the 10 individuals. To do so, we note the following average statistical properties of these observed X_t scores:

1. They are the addition of $n(=5)$ test-sample scores.

TABLE 1

ILLUSTRATIVE SCORE MATRIX, AND THE RELIABILITY COEFFICIENT, r_{tt}, CALCULATED BY FOUR ALTERNATIVE COMPUTING FORMS

Score matrix: $N=10$, $n=5$

	Test-sample, X_i						
Ind.	X_1	X_2	X_3	X_4	X_5	X_t	V_{e_i}
1	6	2	1	0	0	9	4.96
2	8	6	5	2	4	25	4.00
3	10	12	7	7	7	43	4.24
4	5	11	11	9	8	44	4.96
5	6	3	0	0	1	10	5.20
6	11	7	9	6	1	34	11.36
7	7	7	2	5	5	26	3.36
8	4	7	4	4	1	20	3.60
9	6	3	3	2	4	18	1.84
10	6	5	1	3	1	16	4.16
$\sum X_i$	69	63	43	38	32	245	47.68

Individual Variance Form:

$$r_{tt}=1-\frac{\frac{1}{n-1}\left(n^2\overline{V}_{e_i}+M_t^2-n\sum M_i^2\right)}{V_t}$$

$$=1-\frac{\frac{1}{4}[25(4.768)+600.25-5(130.47)]}{140.05}$$

$$=1-.120$$

$$=.880$$

$$\overline{V}_{e_i}=47.68/10=4.768$$

Variance Form:

	X_1	X_2	X_3	X_4	X_5	X_t	$\sum_i$
M	6.90	6.30	4.30	3.80	3.20	24.50	24.50
M^2	47.61	39.69	18.49	14.44	10.24	600.25	$130.47=\sum M_i^2$
$V=\sigma^2$	4.29	9.81	12.21	7.96	7.16	140.05	$41.43=\sum V_i$
						$=V_t$	$\overline{V}_i=8.286$

$$r_{tt}=\frac{n}{n-1}\left(1-\frac{\sum V_i}{V_t}\right)$$

$$=\frac{5}{4}\left(1-\frac{41.43}{140.05}\right)$$

$$=.880$$

Part-Whole Form:

	X_1	X_2	X_3	X_4	X_5	
$\sum X_iX_t$	1804	1892	1427	1245	1035	
r_{it}	.463	.940	.903	.941	.793	
$\sigma_i r_{it}$	.959	2.945	3.156	2.653	2.121	$\sum \sigma_i r_{it}$ $=11.834$

$$r_{tt}=\frac{n}{n-1}\left[1-\frac{\sum V_i}{(\sum \sigma_i r_{it})^2}\right]$$

$$=\frac{5}{4}\left(1-\frac{41.43}{11.834^2}\right)$$

$$=.880$$

Covariance (and its Approx.) Form (Variance-covariance matrix*):

	X_1	1.000 4.29	.313 2.03	.377 2.73	.270 1.58	.130 .72
	X_2	.313 2.03	1.000 9.81	.778 8.51	.923 8.16	.756 6.34
X_j	X_3	.377 2.73	.778 8.51	1.000 12.21	.848 8.36	.593 5.54
	X_4	.270 1.58	.923 8.16	.848 8.36	1.000 7.96	.707 5.34
	X_5	.130 .72	.756 6.34	.593 5.54	.707 5.34	1.000 7.16

Covariance Form:

$$r_{tt}=\frac{n\bar{c}_{ij}}{\overline{V}_i+(n-1)\bar{c}_{ij}}$$

$$=\frac{5(4.931)}{8.286+4(4.931)}$$

$$=.880$$

Covariance Approx. (Spearman-Brown) Form:

$$r_{tt}\doteq\frac{n\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}}$$

$$\doteq\frac{5(.5695)}{1+4(.5695)}=.869$$

2. The *mean variance*, $\overline{V}_i$, of the 5 constituent test-samples is $\overline{V}_i = \overline{\sigma_i^2}$ (=8.3, see 4th line from the bottom).

3. The *mean covariance*, $\bar{c}_{ij}$, between the 5 test-samples is $\bar{c}_{ij} = \sigma_i\sigma_j r_{ij}$ (=4.9, see 3rd line from the bottom).

To compute the reliability coefficient, r_{tt}, we conceptualize a second composite, X_t', whose constituent test-sample scores of the 10 individuals may take *any* values subject only to the following defined conditions:

1. There be $n(=5)$ of them.

2. Their *mean variance*, $\overline{V}_{i'}$, equal that in the observed matrix, that is, $\overline{V}_{i'} = \overline{V}_i(=8.3)$.

3. Their *mean covariance*, $\bar{c}_{i'j'}$, equal that in the observed score matrix, i.e., $\bar{c}_{i'j'} = \bar{c}_{ij}(=4.9)$.

4. The *mean cross covariances* between the test-samples of X_t and those of X_t' preserve certain relations to one another depending on the *structure* of X_t, whether its test-samples are unstratified or stratified:

Unstratified composites: If the observed test-samples are not ordered or grouped in any known way but are as if drawn at random from a large pool of test-samples, then by definition the test-samples of the construct composite, X_t', must be similarly composed, hence the mean cross covariance, $\bar{c}_{ij'}$, would equal the observed mean covariance, $\bar{c}_{ij}$.

Stratified composites: If the observed test-samples are, however, ordered or grouped by known strata, then by definition so must be those of the comparable construct. We will consider this type of structure in a later section.

Unstratified Composites and Domains

Under the definition of an unstratified comparable construct, X_t', we can compute *exactly* the reliability coefficient, r_{tt}, from the observed constants of X_t and without further restrictive conditions.

Let us now list, generally, the defined statistical properties of a comparable construct composite, X_t', in terms of the observed values of X_t:

$$n' = n \qquad [7]$$

(Equality of number of test-samples),

$$\overline{V}_{i'} = \overline{V}_i \qquad [8]$$

(Equality of mean variance),

$$\bar{c}_{i'j'} = \bar{c}_{ij'} = \bar{c}_{ij} \qquad 9]$$

(Equality of mean covariances).

It follows from these definitions that the variance, $V_{t'}$, of the second composite equals the observed V_t since in the formula for the variance of a sum (**5**, p. 586) their parallel terms are equal by [7], [8], and [9], i.e.,

$$V_t' = n'\overline{V}_{i'} + n'(n'-1)\bar{c}_{i'j'}$$
$$= n\overline{V}_i + n(n-1)\bar{c}_{ij} = V_t. \qquad [10]$$

General Form for r_{tt}. The correlation, r_{tt}, is simply the Pearson r between the sum, X_t, defined by [1], and the sum, X_t' by [2]. By the formula for the correlation between sums (**5**, p. 597)

$$r_{tt} = \frac{(1/N)\sum(x_1 + x_2 + \cdots + x_n)(x_1' + x_2' + \cdots + x_n')}{\sigma_t\sigma_t'}.$$

[*] 1st entry is r_{ij}; 2nd entry is $\sigma_i\sigma_j r_{ij} = c_{ij}$
2nd entries:
$\sum$ diagonal $= \sum V_i$ = 41.43; $\overline{V}_i = 41.43/5 = 8.286$
$\sum$ remainder $= 2\sum c_{ij}$ = 98.62; $\bar{c}_{ij} = 98.62/20 = 4.931$
$\sum$ all $= V_t$ = 140.05
1st entries: $\sum r_{ij} = 5.695$; $\bar{r}_{ij} = 5.695/10 = .5695$

The numerator consists of the cross covariances, n^2 in number, between the test-samples of X_t and X_t', and hence reduces to $n^2\bar{c}_{ij}$ by [9]. The σs in the denominator are equal from [10]. The general formula reduces, then, to

$$r_{tt}=\frac{n^2\bar{c}_{ij}}{V_t} \qquad [11]$$

(*General Form* of the reliability of an unstratified composite).

Alternative computing formulas for r_{tt}. To calculate this one value one may, however, use any one of four computing formulas which differ not in the answer they give but in certain constants of the score matrix which the analyst may prefer to use. These computing forms are:

The Variance Form, variously called *Alpha*, or L_3, or for dichotomous variables the Kuder-Richardson (or K-R) formula 20.

The Part-Whole Form, a special case of which is called "Gulliksen's formula."

The Individual Variance Form, not reported elsewhere to the writer's knowledge.

The Covariance Form, an approximation to which is known as the Spearman-Brown formula.

Confusion about these computing forms has been due to the fact that different writers have derived special cases or approximations of them. In their general forms they are *identities*, as shown in Table 1 where they all have the same value of $r_{tt}=.880$. Further confusion has arisen because different writers have derived them on the basis of different assumptions or restrictions, thus leaving their readers in doubt about their application to real data. We shall see that no conditions other than the definitions given above in [7], [8], and [9] are necessary. The four computing forms evaluate the general form, [11], by substituting in it terms easier to compute. Let us examine them in detail below.

The Variance Form (*Alpha*, L_3, *and the K-R special case*). The simplest way to evaluate [11] is to solve for $\bar{c}_{ij}$ in [10], and substitute its equivalent in [11], whence

$$\begin{aligned} r_{tt} &= \frac{n^2}{V_t}\left[\frac{V_t-n\bar{V}_i}{n(n-1)}\right] \\ &= \frac{n}{n-1}\left(1-\frac{\sum V_i}{V_t}\right) \qquad [12] \end{aligned}$$

(*Variance Form: Reliability from variances of test-samples and* X_t).

The computations are illustrated in Table 1 under "Variance Form" where the sequence of desk calculator operations for M, M^2, and V both of the test-samples and of total X_t scores is shown. Substituting the appropriate values in the Variance Form at the right gives $r_{tt}=.880$.

The Variance Form is called *Alpha* by Cronbach (**4**), L_3 by Guttman (**8**), and the Kuder-Richardson Formula 20 (**16**) for the special case of dichotomous items.

The Part-Whole Form ("*Gulliksen's formula*"). For purposes of item analysis, one may be interested in the relation between each test-sample and the total composite score. He would calculate the correlation, r_{it}, between each test-sample, X_i, and the total score, X_t. These correlations may then be used directly to compute the reliability, thus:

$$r_{tt}=\frac{n}{n-1}\left[1-\frac{\sum V_i}{(\sum \sigma_i r_{it})^2}\right] \qquad [13]$$

(*Part-Whole Form: Reliability from r's between test-samples and* X_t).

Table 1 under "Part-Whole Form" gives the requisite desk calculator operations. Note that the value of r_{tt} is exactly .880. The several values of $\sigma_i r_{it}$ are of interest since they reveal the relative contribution of the different test-samples to the reliability of the composite.

Gulliksen derived the Part-Whole Form under restrictive assumptions for the special case of dichotomous items (**7**, p. 378), but you can see that the form is general because the Part-Whole Form merely substitutes for V_t in Variance Form [12] the equivalent expression, $(\sum\sigma_i r_{it})^2$. For it can be shown that for any test-sample, X_i, its part-whole correlation with X_t is

$$r_{it}=\frac{V_i+\sum\sigma_i\sigma_j r_{ij}}{\sigma_i\sigma_t} \quad (i\neq j)$$

$$\sigma_i\sigma_t r_{it}=V_i+\sum\sigma_i\sigma_j r_{ij} \quad (i\neq j).$$

If we sum the n such terms for all X_is, the total is V_t by [10], as follows:

$$\sum\sigma_i\sigma_t r_{it}=\sum V_i+\sum\sum\sigma_i\sigma_j r_{ij}=\sigma_t^2$$

$$\sum\sigma_i r_{it}=\sigma_t.$$

$$(\sum\sigma_i r_{it})^2=V_t.$$

The Part-Whole Form is useful as a check. In order to compute the sundry r_{it} values, one must perforce calculate both V_i and V_t. Since these values are, however, the only ones needed in the simpler Variance Form [12] the extra labor of working out the r_{it} correlations is unnecessary.

The Individual Variance Form. If the analyst wishes to study each individual's variance, V_{o_i}, of its n test-sample scores, he can calculate r_{tt} from these individual variances. The formula is

$$r_{tt}=1-\frac{\frac{1}{n-1}\left(n^2\bar{V}_{o_i}+M_t^2-n\sum M_i^2\right)}{V_t} \qquad [14]$$

(*Individual Variance Form: Reliability from individual variances of test-sample scores*).

At the right in Table 1 you see under "Individual Variance Form" that the value of r_{tt} from [14] is also exactly .880. The derivation of [14] from the Variance Form is a little cumbersome, and has been placed in the appendix. We shall see later that the numerator term in [14] is also the value of $\bar{V}_{o_i}$, which is the individual variance of the *total* X_t composite score (see 24a). For another approach to this problem, see Horst (**10**).

The Covariance Form (and the Spearman-Brown approximation). For a cluster or factor analysis of the test-samples the analyst may compute all the intercorrelations between the test-samples. If so, he can calculate r_{tt} from these rs. He would write these rs as the first entries in the *variance-covariance matrix* (or pooling square), as illustrated under "Covariance (and its Approximation) Form" at the bottom of Table 1. If he then multiplies each r by its $\sigma_i\sigma_j$ for the cell, he enters the covariance, c_{ij}, as the second entry. The mean of these covariances, $\bar{c}_{ij}$, may now be used to calculate r_{tt} by the Covariance Form, which is

$$r_{tt}=\frac{n\bar{c}_{ij}}{\bar{V}_i+(n-1)\bar{c}_{ij}} \qquad [15]$$

(*Covariance Form: Reliability from covariances between test-samples*).

Notice that the Covariance Form is simply the General Form [11] in which we have substituted the equivalent of V_t from [10] and cancelled n.

In Table 1, the essential summations for $\bar{V}_i$ and for $\bar{c}_{ij}$ are given below the variance-covariance matrix. The Covariance Form at the right gives $r_{tt} = .880$, as before.

To the writer's knowledge the Covariance Form has not been published as a computing form of r_{tt}. But it is one. Probably it has been ignored because it leads directly to the ***approximation*** familiarly known as the Spearman-Brown formula.

Covariance Approximation Form, or the Spearman-Brown Formula. We can get rid of the covariances in [15] by taking the product of $\bar{V}_i$ and $\bar{r}_{ij}$ as an approximation to the mean covariance, $\bar{c}_{ij}$, i.e.,

$$\bar{c}_{ij} = \overline{\sigma_i \sigma_j r_{ij}} \doteq \bar{V}_i \bar{r}_{ij}. \qquad [16]$$

Substituting [16] in the Covariance Form [15] gives the Spearman-Brown (S-B) formula,

$$r_{tt} \doteq \frac{n\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}} \qquad [17]$$

(***Covariance Approximation Form: the Spearman-Brown formula***).

In Table 1 the evaluation of the S-B formula does ***not*** give the exact value of .880; because of the approximation in [16], it takes the value, .869.

As usually written in texts, the S-B formula is shown with r not as a mean but as a single value on the equivalence assumption [6]. The above development shows that the mean $\bar{r}$ between the test-samples is called for, and that ***no*** assumption of equivalent inter-rs is necessary.

The special fame of the S-B formula stems from the belief that it saves the analyst work in computing r_{tt}. Before assessing this claim, let us recall that the Variance Form [12] requires only the computing of variances.

The orthodox use of the S-B formula is with the ***split-half method***. Here, the test-samples are divided into two halves, X_{h1} and X_{h2}. The total score is thus expressed:

$$X_t = X_{h_1} + X_{h_2}'$$

The Covariance Form [15] and the S-B Form [17] become in this case, respectively,

$$r_{tt} = \frac{2c_{h_1h_2}}{V_h + c_{h_1h_2}}, \qquad [18a]$$

$$\doteq \frac{2r_{h_1h_2}}{1+r_{h_1h_2}}. \qquad [18b]$$

This split-half procedure requires extra work of the analyst who must compute a ***second*** score matrix consisting of the two split-half scores of the individuals, and then he must compute the correlation between these scores. As will be shown later, when the split is odd vs. even test-samples, the use of [18a] or [18b] is desirable when the ***order*** of the test-samples affects their statistical constants.

At best, the answer by the S-B formula is an approximation which should be avoided by the use of the Covariance Form [18a]. For this case of unstratified composites there is the further trouble that the coefficient from [18a] or [18b] is based on only one splitting, and there are of course many arbitrary ways to split-half the composite—random halves, item-parallel halves, odd-even halves—all yielding different values of r_{tt} by [18a] or [18b]. However, the ***mean*** of all these possible split-half rs turns out to be exactly the value given by our Variance Form [12], according to Cronbach (**4**).

Behavior domain validity. A statistic that is more meaningful than the reliability coefficient is the correlation of X_t with a score on a ***domain*** of composites comparable to X_t. A domain score of an individual would be the best criterion of his status in the property, X, as operationally defined. The domain score, usually called a "true" score, is defined as the sum (or average) of scores on a large number of composites, X_t', X_t'', etc., each of which has the same average statistical properties defined in [7], [8], and [9], that is,

$$X_{t_\infty}=X_t+X_t'+X_t''+\cdots+X_{t_{n_\infty}}, \quad [19]$$

where n_∞ is such a large number that

$$1/n_\infty \text{ approaches } 0;\ n/n_\infty \text{ and}$$
$$(n-1)/n_\infty \text{ approach } 1. \quad [20]$$

Such a domain of composite scores is of course a theoretical construct, but an important one, for if the analyst discovers that the correlation, r_{tt_∞}, of the observed composite X_t with such a hypothetical criterion is very high he knows that the individuals are actually ranked in observed scores close to their ranking in a perfectly reliable measure of the property X, as operationally defined. If r_{tt_∞} is low, he knows he must improve his observed sampling of X.

From the correlation of sums, and recalling the defined statistical properties of the comparable composites, X_t', X_t'', $\cdots$ as given in [7], [8], and [9], then

$$r_{tt_\infty}=\frac{(1/N)\sum t(t+t'+t''+\cdots t_{n_\infty})}{\sigma_t\sigma_{t_\infty}}$$
$$=\frac{\sigma_t^2+(n_\infty-1)\sigma_t^2\bar{r}_{tt}}{\sigma_t\sqrt{n_\infty\sigma_t^2+n_\infty(n_\infty-1)\sigma_t^2\bar{r}_{tt}}}\,.$$

Cancelling σ_t^2, dividing numerator and denominator by n_∞, and noting the limits of [20], we get

$$r_{tt_\infty}=\sqrt{r_{tt}} \quad [21]$$

(***Behavior domain validity of X_t***).

This correlation r_{tt_∞}, has been labelled in textbooks as the "index of reliability," a meaningless term. It is obviously the ***behavior domain validity*** of X_t, because it is the correlation between a sample and its perfect criterion measure of the property X as operationally defined. For the data of Table 1, we find that the domain validity of the X_t scores is $\sqrt{.880}$, or the high value of .94.

Individual variance or "error variance." A very practical experimental matter is the discovery of the degree to which an individual's rank in the group of observed X_t scores would probably deviate from his ranking in the true X_{t_∞} scores. Thus in Table 1 we would ask: with what confidence can we believe that individual No. 1, the low scorer with an $X_t=9$, would still be lowest in domain score?

We do not possess the domain score, of course, but we can nevertheless estimate a probable *range* of an individual's observed score around it, at a chosen level of confidence. To do so, let us express the domain score on the scale of the observed scores by writing it as the *average* of the individual's composite scores in [19], i.e.,

$$\overline{X}_{t_\infty}=(1/n_\infty)(X_t+X_t'+\cdots+X_{t_{n_\infty}}). \quad [22]$$

The deviation of an individual's observed X_t score from his domain score, $\overline{X}_{t_\infty}$, is commonly called his "error of measurement" in X_t. This is a bad term because the experimenter usually has no objective grounds for establishing that the fluctuations of an individual's observed performances are "errors"—in fact, they usually are genuine variations that simply deviate from an average "true" parameter value, $\overline{X}_{t_\infty}$, of the individual.

The meaning and importance of assessing this deviation, desirable in all psychological testing, is particularly evident in the case of behavior genetic experiments. How different must the scores of two individuals be in order to be sure they are reliably different? In selective breeding for a pure line one would breed together only those extreme individuals whose scores are *not* reliably different. The experimenter would also cease selective breeding in a "pure" line whose variance in observed scores equalled the individual or "error" variance.

This *individual variance*, called V_{e_t}, is quite simply assessed. It is the variance in observed X_t scores of a subgroup of individuals that are identical in their domain $\overline{X}_{t_\infty}$ scores. Since we know the correlation between X_t and $\overline{X}_{t_\infty}$ from [21], we compute V_{e_t} from the orthodox formula for the standard error of estimate of X_t scores from $\overline{X}_{t_\infty}$ scores, i.e.,

$$V_{e_t}=V_t(1-r_{tt_\infty}{}^2)=V_t(1-r_{tt}), \qquad [23a]$$

whence, σ_{e_t} the *individual standard deviation*, is

$$\sigma_{e_t}=\sigma_t\sqrt{1-r_{tt}}. \qquad [23b]$$

One does not assume that all individuals have the same value of the individual variance. The value by [23a] is the average of the N individual variances, being the *mean* squared deviation around the theoretical regression line of X_t on X_{t_∞}.

To illustrate from Table 1, let us understand the X_t scores there to be the sum of maze errors of rats in five trials (as indeed these scores actually are). The variance between individuals in these scores is $V_t=140$. Were we to selectively breed together over many generations rats with the lowest scores, like individual No. 1, we would cease selection when the line had a variance of $V_{e_t}=140$ $(1-.88)=17$ errors. Further selection would be fruitless since all observed variation between rats in the line would be variance within individual rats and not between them.

A complementary "basic definition" of the reliability. We can express rather meaningfully the reliability coefficient, r_{tt}, as a function of the within-individual variance, V_{e_t}. From [23a] we can write r_{tt} as

$$r_{tt}=1-\frac{V_{e_t}}{V_t} \qquad [24a]$$

(*Individual Variance Ratio Form: Reliability from V_{e_t} of total score, X_t*).

To illustrate from the data of Table 1, the reliability becomes $r_{tt}=1$ $-17/140=1-.12=.88$, as in the other computing forms.

The *individual variance ratio*, V_{e_t}/V_t in [24a], is thus the proportion of the total variance, V_t, among all individuals that is determined by the individual variance, V_{e_t}. Here the ratio is 12%. The complement of this proportion is the value of the reliability coefficient itself which may be written

$$r_{tt}=\frac{V_t-V_{e_t}}{V_t}=\frac{V_{t_\infty}}{V_t}\,. \qquad [24b]$$

Thus the reliability coefficient is that remaining proportion of the total variance which would be the variance of the domain scores of the individuals.

Expressions [24a] and [24b] have been sometimes referred to as the "basic definition" of the reliability coefficient because of the meaningfulness of the proportional terms. You will note that the Individual Variance Ratio Form in [24a] is a parallel identity with the Individual Variance Form [14]. Thus the value of the individual variance, V_{e_t}, of the total X_t score [see 23a] can be computed directly from the mean individual variances, $\overline{V}_{a_i}$ across the test-samples, via the numerator term of [14].

Stratified Composites and Domains

Item-Parallel, Split-Half, Test-Retest and Battery Reliability

In the foregoing treatment we have been examining the unstratified case in which the n test-samples, X_1, $X_2, \cdots X_n$, are drawn with equal likelihood, or without bias, from a potential pool of such test-samples. Situations often arise, however, in which the analyst conceptualizes the property X he seeks to measure as definitely made up of *substrata* of properties. For example, where X is vocabulary knowledge, he may conceptualize it as made up of strata of different levels of difficulty, or of different kinds of content. If X is a personality attribute, the sample judges may come from strata pertaining to degree of acquaintance with the subject. Or the test-samples may be associated with different stages of learning or of test-adaptation of the subjects.

Composites of n strata. "Item-parallel tests." A special case of a composite drawn from a stratified domain is the one in which each of the n items is expressly drawn from one of n defined strata. To understand this case, return to the definitions of the X_t composite in expression [1] and of its comparable construct, X_t', in equation [2]. In this case, X_1 is parallel to X_1', i.e., they are drawn from a defined stratum, like two true-false items of the same kind of content. Similarly, X_2 is parallel to X_2', and X_3 is parallel to X_3', and so on.

In our earlier definition of the comparable construct, X_t', the fourth condition referred to the cross covariances between the test-samples of X_t and its X_t' construct. When, as in this case, the n test-samples fall into n strata, then the mean cross-covariance term of r by [11] is the weighted mean of the following two types of mean cross covariances:

$\bar{c}_{ii'}$, the mean cross covariance between the n pairs of parallel test-samples—an unknown value.

$\bar{c}_{ij'}$, the mean cross covariance between the nonparallel items, n^2-n in number, where $i\neq j'$, and by definition equal to $\bar{c}_{ij}$.

We therefore rewrite the General Form [11] of r_{tt} for unstratified composites in the following form for stratified composites:

$$r_{tt}=\frac{n\bar{c}_{ii'}+(n^2-n)\bar{c}_{ij}}{V_t}\quad(i\neq j)\qquad[25]$$

(*Reliability of a composite with n strata of test-samples*).

To see this form more simply let us take the approximation of [16], substitute the equivalent of V_t from [10], cancel n and $\overline{V}_i$, whence

$$r_{tt}\doteq\frac{\bar{r}_{ii'}+(n-1)\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}}\quad(i\neq j)\qquad[26]$$

(*Approximation to* 25).

The reliability coefficient of a stratified composite by [26] is the analogue of that of an unstratified composite by the familiar S-B formula. The mean correlation, $\bar{r}_{ii'}$, between parallel test-samples will usually be higher than that between nonparallel, i.e.,

$$\bar{r}_{ii'}>\bar{r}_{ij}\qquad[27]$$

whence the reliability of composites of n strata will usually be higher than that for unstratified composites. The lowest value that $\bar{r}_{ii'}$ could reasonably take in any composite would be $\bar{r}_{ii'}=\bar{r}_{ij'}$. Substitution of this lowest limit in [25] and [26] reduces them, respectively, to the General Form [11] and the Spearman-Brown formula [17] for unstratified samples. Therefore, when one is dealing with a stratified composite the lower limit of its reliability is that found by any of the earlier computing forms for an unstratified composite.

The troublesome feature of an observed composite made up of n strata is that its reliability by [25] or [26] is indeterminate without a second set of n parallel test-samples from which $r_{ii'}$ can be calculated. Without such a comparable set of test-samples one must usually be content with finding the lower limit of r_{tt} from [11].

Odd-even split-halves. When the n strata of X_t represent test-samples that are a serial order of responses of the subjects, usually the case with psychological tests and learning tests, a solution of r_{tt} is available by the odd-even split-half method. By this procedure one scores each subject on two subcomposites as follows:

$$X_{h_o}=X_1+X_3+X_5$$

and so on for all odd test-samples,

$$X_{h_e}=X_2+X_4+X_6$$

and so on for all even test-samples.

Here one has two comparable composites which would satisfy the four defined conditions of comparability including the approximate matching of test-samples for serial order. The correlation, $r_{h_oh_e}$, is thus by definition the reliability of a composite with one-half the number of test-samples but with the same serial stratification of the total composite, X_t. Substituting $r_{h_oh_e}$ into [18a] or [18b] gives the desired value of r_{tt}.

This common method of determining the reliability of a composite is the correct procedure to follow in all those testing situations where the test-samples do have a serial order. It should give values for reliabilities either equal to but usually higher than that computed by [11] for unstratified composites—a fact commonly observed (e.g., **12**, Ch. 4).

Replicated composites ("*Test-retest reliability*"). The analyst who wishes to observe the constancy of individual differences over time may administer the same composite, X_t, on several occasions. In this special case, the actual replications, X_t', X_t'', are *duplicate* parallel composites, test-sample for test-sample, rather than matched test-samples, as just above.

Before actually performing these replications, the analyst can estimate the lower and upper limits of the correlations between the observed X_t composite and the proposed replications of it. These limits can be estimated from the constants of one administration of X_t alone. At the lower limit the covariance, $\bar{c}_{ii'}$, between duplicate test-samples on replication would probably not be lower than the covariance, $\bar{c}_{ij}$, between nonparallel test-samples on the first occasion (if the proposed time interval is reasonably short), hence for $\bar{c}_{ii'}=\bar{c}_{ij}$ in [25], we find the probable lower limit of the test-retest correlation to be that between unstratified composites by [11]. There is, however, no prior knowledge of the effect of elapsed time on the property, X. It could easily be that between the first administration of X_t and its second, X_t', the cross covariances, $\bar{c}_{ii'}$ and $\bar{c}_{ij'}$, may be much less than $\bar{c}_{ij}$, and in the limit it is possible that the test-retest coefficient could go to zero, or negative.

At the upper limit, the correlation between replicated test-samples could be 1.00, and their variances could remain constant over time. In this case, you will note that in [25], and more obviously in [26], the numerators would equal the denominators, whence the upper limit of the test-retest coefficient would be 1.00. Guttman (8) has studied these limits in some detail, basing them, however, on certain restrictive assumptions not required in the above analysis.

Composite of test-samples with variable N (*"Speed tests"*). A "speed test" with a restricted time limit may reveal an increasing proportion of the N subjects falling into the class, "No response," on the n successive test-samples, X_1, X_2, $\cdots$ X_n; similarly, with a "power test" having test-samples of increasing difficulty. The reliability, r_{tt}, by [26] becomes indeterminate in this case because it is not possible to compute either $\bar{r}_{ii'}$, or $\bar{r}_{ij}$, even if one had at hand an actual comparable composite with parallel test-samples X_1', X_2', $\cdots$, X_n'. The reason is that "No response" does not fall in the same continuum with the quantitative scores of those who do respond. Our formulations only apply to that fraction of all N subjects who respond to all n test-samples—usually a subgroup of restricted range.

A special experimental design is required of a test if it is to satisfy the definition of a speed test: all subjects should respond to each of the n test-samples, the measure taken on them being *elapsed time*. Under such a design, reliability or its limits becomes determinable by the formulations presented here. With a graded power test our formulations also apply provided the class "No response" can be legitimately assigned the lowest score possible on the test-sample, usually zero.

Subdomains (*Battery reliability*). The more general case of a stratified domain, and one which does provide an exact solution of the reliability coefficient from the constants of the observed composite, is the one in which the strata represent defined subdomains of test-samples. Common examples would be a vocabulary test with one stratum being a block of items at one difficulty level, a second stratum of items at another level, and so on. Test "batteries," such as the Wechsler-Bellevue, are of this form. The reliability coefficient of the grand total score, X_t, is calculated either from odd-even split-halves or, as shown below, from constants of the blocks of observed test-samples.

In this general case, the composite scores of X_t and of another comparable construct composite, X_t', as defined by [1] and [2] may be written:

$$\left.\begin{aligned} X_t &= \sum X_g \\ &\quad + \sum X_h + \cdots + \sum X_k \\ X_t' &= \sum X_g' \\ &\quad + \sum X_{h'} + \cdots + \sum X_k'. \end{aligned}\right\} \qquad [28]$$

where the parallelism is of the subdomains, g, h, $\cdots$, k, with parallel blocks of test-samples $\sum X_g$ and $\sum X_g'$, $\sum X_h$ and $\sum X_h'$, and so on, there being k such parallel blocks of test-samples. By our definition of comparability, there are n_g test-samples in $\sum X_g$, n_h in $\sum X_h$, and so on. For parallel blocks, equal mean variances and covariances of conditions of [8] and [9] hold by definition.

From the correlation of sums, the reliability of X_t, which is the correlation between X_t and X_t' of [28] becomes, after a little algebra,

$$r_{tt} = \frac{\sum r_{t_g t_g} V_{t_g} + 2 \sum c_{t_g t_h}}{V_t} \qquad [29]$$

(*Covariance Form of the reliability of a stratified composite*), where $r_{t_g t_g}$ is the reliability of the total score on each of the k sets of X_g block of test-samples, simply computed by the Variance Form [12], V_{t_g} is the computed variance of the total scores on the X_g block; $\sum c_{t_g t_h}$ is the sum of covariances between total scores on different X_g and X_h $(g \neq h)$ blocks; V_t is the variance of the grand total scores on the stratified composite, X_t, calculated from the variance of a sum, i.e.,

$$V_t = \sum V_{t_g} + 2 \sum c_{t_g t_h}. \qquad [30]$$

A simpler computing form results from solving for $\sum c_{t_g t_h}$ in [30], and substituting its equivalent in [29], whence we get the identity

$$r_{tt} = 1 - \frac{\sum V_{t_g} - \sum V_{t_g} r_{t_g t_g}}{V_t} \qquad [31]$$

(*Variance Form of the reliability of a stratified composite*).

For the relatively common situation in which the total scores on the different strata are in different units, the analyst would convert them to sigma scores, with $V_{t_g} = 1$, whence [31] reduces to

$$r_{tt} = 1 - \frac{n_t - \sum r_{t_g t_g}}{V_t'} \qquad [32]$$

(*Variance Form of* 31 *for equally weighted strata*),

where n_t is the number of subdomains or strata, V_t' is the variance of the equally weighted total score.

There are no unknowns in this situation, and hence an exact value of r_{tt} for such a stratified composite can be calculated. In Formula 29 or 31 we have a *completely general* formula for a determinable reliability coefficient, r_{tt}. For the special case of a stratified composite made up of n parallel test-samples, treated in the preceding two sections, Formula 29 reduces to Formula 25. For the case of unstratified domains, the numerator of [29] reduces to $n^2 \bar{c}_{ij}$ which leads to our General Form [11], whose approximation is the orthodox Spearman-Brown formula [17].

Simplified Formulation for Composites of Dichotomous Test-Samples (Items)

Part of the confusion about the different forms of computing the reliability of a composite is due to the fact that, since psychologists have focused on mental tests or questionnaires composed of dichotomous items, like Yes-No questions, some of the computing forms for reliability have been publicized by their special cases for this situation, like the K-R formula, while others are known by their general case for continuous test-samples, like the Variance Form (*Alpha*) or the S-B formula.

When the test-samples are dichotomous, the score matrix has 1's and 0's in it instead of continuous values as in Table 1. In such a special case, $V_i = p_i q_i$, where p_i is the proportion of Yes's or 1's in test-sample X_i, and $q_i = 1 - p_i$. For this special case we therefore insert $\sum p_i q_i$ for $\sum V_i$ in Variance Form [12]. This computational variant of the Variance Form has been labelled the K-R Formula 20 after its authors (**16**). This appellation is undesirable not only because the K-R form is simply a minor situational variant of the Variance Form but primarily because its derivers assumed that the test-samples are determined by one general factor and should have equal variances and inter-rs. The direct derivation of [12] from [11] demonstrates that these "assumptions" are, in fact, unnecessary restrictions.

Similarly, the Part-Whole Form

[13] was derived by Gulliksen for the purely Yes-No test-samples. Here, one needs the Pearson r of a dichotomous item with total score, a continuous variable. The computational form of r simplifies in this case to the *point-biserial* r, but the Part-Whole Form by [13] is in no way restricted to point-biserials, as our derivation shows.

In like fashion, an analyst using the Covariance Form [15] or its S-B approximation [16] which require the inter-rs between test-samples, would compute these rs from the *phi coefficient*, the product-moment r between dichotomous test-samples.

For the Individual Variance Form [14], one would need to calculate each individual's variance, $V_{o_i} = p_o q_o$, where p_o is simply $\sum(\text{Yes's})/n$, and $q_o = 1 - p_o$.

The Total Score Form. The speediest means of determining the reliability of a dichotomous-item test is by use of the Total Score Form which requires only the mean and variance of the total X_t scores. This approximation is in fact the *only* means of determining the reliability in experimental situations where the analyst may not be able or may not wish to record the performance of subjects on the n test samples but records only their final total X_t scores. An example is Drosophila experiments where the flies physically traverse the n test-sample situations but ultimately wind up in test tubes corresponding to the different classes of the X_t total score scale (**9**).

The Total Score Form for computing r_{tt} derives from the Variance Form [12] in which the essential terms are $\sum V_i = \sum p_i q_i$, and V_t. The latter term, V_t, is, of course, computed directly from the distribution of total X_t scores. We can approximate the $\sum p_i q_i$ term (dropping the i subscript) by the development below:

$$\sum pq = \sum (p - p^2) = \sum p - \sum p^2$$
$$= n(\overline{p}) - n(\overline{p^2}). \qquad [33]$$

From the formula for the mean of a sum, we note that

$$n(\overline{p}) = M_t \qquad [34]$$

Since the p values of the different items are not recorded we cannot exactly compute the mean of the squared ps, i.e., $\overline{p^2}$. As an approximation we can take the mean of their squares to be the square of their mean, i.e.,

$$\overline{p^2} \doteq \overline{p},^2 \qquad [35]$$

which will be close if the p values of the items do not vary too greatly. We can write the approximation as follows:

$$n(\overline{p^2}) \doteq n(\overline{p})^2 = M_t^2/n. \qquad [36]$$

Substituting [34] and [36] in [33], and writing the Variance Form [12] in full,

$$r_{tt} = \frac{n}{n-1}\left(1 - \frac{M_t - M_t^2/n}{V_t}\right) \qquad [37]$$

(*Total Score Form: Approx. reliability from the total X_t scores only*).

The Total Score Form is known as the Kuder-Richardson Formula 21 (**16**). The reliability determined by the use of it is a lower limit of the correct value which would be found by the Variance Form [12] or odd-even split-half. There are no "assumptions" in the derivation of the Total Score Form; it involves only the approximation of [36] and it approaches the correct value according as the item ps approach equality. Experience seems to show that a relatively wide range of p values can be tolerated without the value from [37] deviating greatly from the correct value.

History of Orthodoxy in Mental Test Theory

In the preceding section we have seen that for unstratified composites all the commonly used formulas for determining the reliability coefficient —the Variance Form [12], the Part-Whole Form [13], the Individual Variance Form [14] and the Covariance Form [15]—are identities, being merely computing forms of the General Form [11]. We have also seen that the derivation of them involves no assumptions of "underlying factors," or of statistically equivalent test-samples. They are quite simply derived on the objective principles of domain sampling. The test-samples that make up the composite X_t are taken as they come, with unequal variances and covariances.

Surprising it is, therefore, that virtually all writers on mental test theory over the last half century have clung either to the orthodox theory of true and error factors, or the theory of equivalent test-samples, the first a set of unverified postulates, the second obviously unrealistic. An attempt to explain why psychologists have been unable to free themselves from these two rigid mental sets over about 50 years will not be made here. This section endeavors merely to document this orthodoxy.

We find an early, clear formulatiọn of the truth-error factor theory by Spearman in 1910 (**18**). He expresses test-samples as "$x_1, x_2, \cdots = x+d_1, x+d_2$, where x is the underlying regular measurement, while the *d*s are the superimposed accidental components" (p. 289). From these postulates he then develops his famous formula (our Formula 17), later known as the Spearman-Brown formula. We note here also the expression "underlying" which has come down through time as a key conception of factor analysts.

In the same edition of the *British Journal of Psychology*, William Brown takes a more objective view of test-samples and derives the same S-B formula on the theory of equivalent test-samples (**1**, footnote p. 299; see also **2**). We thus have in these writings the beginnings of the two opposing orthodoxies.

Spearman's truth-error doctrine dominated thinking for a considerable time thereafter, even of the opponents of Spearman's general factor theory. In his general statistics text Yule (**27**) adopts the truth-error formulation of Spearman but with the caution "if the further assumption is legitimate that the errors in d_1 and d_2 are uncorrelated with each other" (p. 212).

In their *Essentials of Mental Measurement* in 1922, Brown and Thomson (**3**) have a real go at the Achilles heel of the truth-error postulate, namely, the belief that the alleged errors of measurement, the *d*s, are "accidental," "random," and "uncorrelated." As they pungently put it, "When an individual [is measured by two test-samples] there is no error of observation involved. [The scores] are the actual true measures of ability on the two occasions. The average or mean ability $[x_\infty]$ $\cdots$ is doubtless different from either, but that does not make the other two measures erroneous. [The deviations of these scores from the mean ability] represent *individual variability*, and to assume them uncorrelated with one another or with the mean values is to indulge in somewhat *a priori* reasoning" (p. 158). However, Brown and Thomson do not themselves otherwise formulate the problem.

It is Kelley in his 1924 *Statistical*

Methods (**13**) who appears first to organize clearly and elegantly the derivation of reliability on the assumption of equivalent test-samples. He begins his development by reducing the test-samples to z scores thus making their variances equal; then further assuming their inter-rs to be equal, he develops all the basic formulas. Just as the Spearman-Yule formulation of the truth-error doctrine sets the orthodoxy on the factorial side, so has Kelley's formulation been the bible for proponents of the equivalent test-sample doctrine.

Soon after, in Holzinger's 1928 *Statistical Methods for Students of Education* (**11**) we find the paradox of a writer's accepting both orthodoxies, but in different parts of his text. In developing the S-B formulation of the reliability coefficient, Holzinger follows Kelley's development (p. 168) but in deriving the standard error of measurement (p. 250 ff.) he strictly follows the Spearman-Yule truth-error postulates.

In 1930, fresh from graduate school, the present writer wrote an article on the reliability coefficient (**21**) following in part the pattern of the truth-error conception. By 1935 he knew better, and in a paper rejecting the whole concept of factors as being psychologically and biologically unrealistic dismissed the notion that individual variability could be thought of as "error" (**22**). He has not budged from this position since, and in place of the current vogue for factor analysis he has developed the methods of cluster analysis (**23**, **24**, **25**) based on the same general principles of domain sampling as expressed in this paper.

The factorial conception of a test-sample score had its modern dress-up by Thurstone. In 1931 he started conventionally with a short brochure on *The Reliability and Validity of Tests* (**19**) where he formulates the problem in orthodox Spearman-Yule fashion. But by 1935 in *The Vectors of Mind* (**20**) he had fully developed the now familiar postulate of breaking up X_∞ into several additive "underlying" multiple factors plus an uncorrelated specific. This conception is, however, a rigid adherence to the truth-error doctrine, though it contributes a new interpretation of "truth."

A different approach to the method of computing reliability is presented in the 1937 publication of the Kuder-Richardson Formula 20 (**16**), which determines the reliability coefficient of a test from the variances of its dichotomous items. Confusion has been considerable over the "assumptions" of this formula because the authors derive it via the truth-error doctrine, alleging that the test-samples must measure only one common factor and also be statistically equivalent. We have seen, however, that the K-R formula is but a special case of the Variance Form [12], and hence involves ***no*** assumptions whatsoever about the factorial composition or equivalence of the test-samples.

The next landmark is the excellent mathematically based 1940 statistics text by Peters and Van Voorhis (**17**). They take no original approach to our problem, however, following the formulations of Kelley and thus becoming faithful followers of the equivalent test-sample camp.

The same year Jackson and Ferguson published their provocative monograph on the reliability of tests (**12**). On the grounds that "implicit in the reliability concept is the idea of repeated measurement" (p. 77), they seem, however, to accept the empirical correlation between comparable forms, split halves, test-retest as better measures of reliability than the correlation between a test and its comparable construct. They turn to variance analysis to separate out "errors of measurement," true individual differences, and practice effects. In their examination of the K-R formula they trenchantly note that, as we have emphasized here, Kuder and Richardson "fell into the common error of specifying conditions that are sufficient but unnecessary" (p. 72), and that the less restrictive condition of equal covariances (our expression 9) is required (p. 75). In their treatment of the reliability of a battery (what we have called a "stratified composite"), they do, however, accept as its reliability its correlation with a comparable construct and derive from it the basic formulas for this situation. Had these writers adopted this same conception of reliability in their treatment of the reliability of an unstratified composite, their formulation of this general problem would not have deviated substantially from that of the present author.

A complete break with orthodoxy is Guttman's 1945 treatment (**8**) of test-retest reliability of a composite. His objective is to establish lower limits of this coefficient from the known constants of the test-samples of X_t. His signal finding is that our Variance Form [12], his L_3, is such a lower limit. However, to achieve such a limit he introduces a *different* set of assumptions from the usual. He says, "We use essentially only one basic assumption; that the errors of observation are independent between items [our test-samples] and between persons over the *universe of trials* [our X_t, X_t', X_t'', $\cdots$]. In the conventional approach, independence is taken over *persons* rather than trials" (p. 257). Now, "independent errors of observation" clearly implies a factorial construction of test-sample performance, and hence is a postulate requiring substantiation. His basic assumption, though interestingly new, is quite unnecessary, for we have seen above in our treatment of the test-retest situation that, without any assumptions, the Variance Form, or Gutman's L_3, is a lower limit of the test-retest coefficient.

An interesting conversion was that of Kelley in his rather monumental 1947 *Fundamentals of Statistics* (**15**). Here he deserts the objective approach which in large part he initiated in 1924, and joins the truth-error votaries. Signs of this conversion are nevertheless implicit even in his earlier 1924 book where he is preoccupied with rules for the construction of two comparable tests designed to cut down correlation between "errors" in order to arrive at the "true reliability coefficient" (p. 201 ff.). After 1924 he became a factor analyst of his own special sort (**14**), and thus by 1947 his concept of a test-sample is that it is either "an expression of [a] common function-plus-chance, or of this common function-plus-a-non-chance-unique-function-plus-chance" (**15**, p. 401). Starting with this orthodox factorial construction, he develops all the essential basic formulas we have developed earlier in this paper.

The most comprehensive attack on the problem is Gulliksen's 1950 *Theory of Mental Tests* (**7**). His formulations are, however, largely restricted to the case of dichotomous test-samples, such as true-false test items. He first presents the truth-error doctrine in orthodox form, unfortunately labeling it "The basic assumption of test theory" (Ch. 2, p. 4), and develops all the basic formulas from it. However, by Chapter 3 he sheds this "basic assumption" and for the rest of the book develops the general and specialized formula for reliability and validity on the postulates of equivalent test-samples, called "parallel tests" by him. These are defined objectively "in terms of observable characteristics," namely, equal means, equal sigmas, equal inter-*r*s (pp. 28-29). When he comes to the treatment of the Kuder-Richardson formula, he nearly breaks free, for he derives the K-R formula without the crippling factorial and other restrictive assumptions of its original authors. However, this derivation is developed for two composite tests, X_t and X_t' "that are parallel item for item" (p. 221), which is the special case developed earlier in this paper under "Item-parallel tests" (see our Formula 26). His derivation of the Variance Form via item-parallel test-samples is most unfortunate, for you may recall that for stratified composites one needs $\bar{c}_{ii'}$, the covariance between parallel items, and $\bar{c}_{ij}$, that between nonparallel. Since an analyst rarely has available the second test, X', the covariance between parallel items is unknown, so in this dilemma, Gulliksen retreats to the orthodoxy of equivalent test-samples and states, unrealistically, "the simplest and most direct assumption is that . . . the covariance between parallel items is equal to the . . . covariance between nonparallel items" (p. 223), an assumption our expression [27] rejects. Had Gulliksen approached the K-R formula, or the more general Variance Form, not for item-parallel tests but for unstratified composites, he would not have run into the $\bar{c}_{ii'}$ term and probably would have seen that for the Variance Form the equivalent test-sample doctrine is an unnecessary restriction of it or of *any* of the basic formula. He also derives the Part-Whole Form, our [13], but to get it he employs the K-R formula as a step, and thus bounds the Part-Whole Form by the same restrictions and the dichotomous item situation he sets for the K-R formula, all quite unnecessary as we have shown.

In 1951 Cronbach's treatment (**4**) of the Variance Form for computing r_{tt}, which he calls *Alpha*, comprehensively explores the usefulness of this computing form. However, Cronbach does not derive the formula in a clear way that permits the reader to see what assumptions, if any, are taken to arrive at it. He shows how to interpret the Variance Form in terms of the truth-error factorial postulate (p. 312). He also states that with respect to the correlation between test-samples, it is "among items having equal variances and equal covariances" (p. 323), the equivalent test-sample postulate. One gains the impression from Cronbach that the Variance Form is a generalized formula, but we have seen that the General Form [11] is more general and that [29] is even more general. The Variance Form shares equivalent status with the Covariance, Part-Whole, and Individual Variance Forms in being another alternative computing form of [11].

We finally come to the 1954 revision of Guilford's *Psychometric Methods* (**6**). This generally excellent reference work is marred for the present writer by the strict adherence of Guilford to the truth-error factor doctrine, which he labels "The Rationale of Test Reliability" (p. 349). All the computing forms for reliability and domain validity of composites are nicely organized by Guilford, but his reader is not let in on the fact that they can be objectively derived even on the equivalent test-sample doctrine. The reader is led to believe that the truth-error doctrine is the *only* mental test rationale. The reader of the foregoing pages will now see that the manifold formulas in test construction, so ably arrayed by Guilford, derive quite directly from the objective principles of measurement by domain sampling, and that the factor postulates of Guilford are just one type of orthodoxy, and quite unnecessary if the reader has no predilection for "underlying" factors.

Appendix

Derivation of the Individual Variance Form. The variance of any individual across the n test-samples is V_{e_i}, where

$$V_{e_i}=\sum X_i^2/n-\left(\sum X_i/n\right)^2.$$

The mean individual variance becomes, noting that $\sum X_i = X_t$,

$$\overline{V}_{e_i}=\frac{\sum}{N}\left[\frac{\sum X_i^2}{n}-\frac{X_t^2}{n^2}\right]$$

$$=\frac{1}{n}\left[\frac{\sum}{N}\left(\sum X_i^2\right)\right]-\frac{1}{n^2}\left(\frac{\sum X_t^2}{N}\right). \qquad [38]$$

The first term of the member on the right can be found from constants of the n test-samples, i.e.,

First term $=(1/n)(\sum V_i+\sum M_i^2)=\overline{V}_i+(1/n)\sum M_i^2$.

The second term on the right $=(1/n^2)(V_t+M_t^2)$.

Substituting these two terms in [38], solving for $\overline{V}_i$, we get after a little algebra,

$$\overline{V}_i=\overline{V}_{e_i}+(1/n^2)V_t+(1/n^2)M_t^2-n\sum M_i^2. \qquad [39]$$

Now, to find r_{tt}, we substitute [39] for $\overline{V}_i$ in the Variance Form in which $\sum V_i=n\overline{V}_i$, multiply numerator and denominator of the fraction by n, then after a little manipulation we finally get the Individual Variance Form [14].

References

1. Brown, W. Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.*, 1910, **3**, 296–322.
2. Brown, W. Some experimental results in correlation. *C. R. VI^me Congr. Int. Psychol.*, Genève, 1910.
3. Brown, W., & Thomson, G. H. *The essentials of mental measurement.* Cambridge: Cambridge Univer. Press, 1921.
4. Cronbach, L. J. Coefficient alpha and the internal structure of test. *Psychometrika*, 1951, **16**, 297–334.
5. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1950.
6. Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.
7. Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.
8. Guttman, L. A basis for analysing test-retest reliability. *Psychometrika*, 1945, **10**, 255–282.
9. Hirsch, J., & Tryon, R. C. Mass screening and reliable individual measurement in the experiment behavior genetics of lower organisms. *Psychol. Bull.*, 1955, **53**, 402–410.
10. Horst, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, **14**, 21–31.
11. Holzinger, K. J. *Statistical methods for students of education.* New York: Ginn, 1928.
12. Jackson, R. W. B., & Ferguson, G. A. *Studies on the reliability of tests.* Bull. 12. Toronto: Dept. Educ. Res., Univer. of Toronto, 1941.
13. Kelley, T. L. *Statistical methods.* New York: Macmillan, 1924.
14. Kelley, T. L. *Crossroads in the mind of man.* Stanford: Stanford Univer. Press, 1928.
15. Kelley, T. L. *Fundamentals of statistics.* Cambridge: Harvard Univer. Press, 1947.
16. Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, **2**, 151–160.
17. Peters, C. C., & Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.
18. Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.*, 1910, **3**, 271–295.
19. Thurstone, L. L. *The reliability and validity of tests.* Ann Arbor: Edwards, 1931.
20. Thurstone, L. L. *The vectors of mind.*

Chicago: Univer. of Chicago Press, 1935.

21. Tryon, R. C. The reliability coefficient as a per cent, with application to correlation between abilities. *Psychol. Rev.*, 1930, **37**, 140–157.

22. Tryon, R. C. A theory of psychological components—an alternative to "mathematical factors." *Psychol. Rev.*, 1935, **42**, 425–454.

23. Tryon, R. C. *Cluster analysis.* Ann Arbor: Edwards, 1939.

24. Tryon, R. C. Identification of social areas by cluster analysis. Berkeley: *Univer. of Calif. Publ. Psychol.*, 1955, **8**, 1–100.

25. Tryon, R. C. General dimensions of individual differences: Cluster analysis vs. factor analysis. Paper read at West Psychol. Ass., Berkeley, March, 1956. On file, Univer. of Calif. Library, Berkeley, Calif.

26. Tryon, R. C. Communality of a variable: Reformulation from cluster analysis. Psychometrika, in press.

27. Yule, G. U. *An introduction to the theory of statistics.* London: Griffin, 1922.

Received September 10, 1956.

Psychological Bulletin
Vol. 54, No. 3, 1957

A TABLE TO FACILITATE COMPARISON OF PROPORTIONS BY SLIDE RULE

ANDREW R. BAGGALEY
University of Wisconsin-Milwaukee

One of the statistical problems most frequently confronting the research worker in psychology is the comparison of proportions or percentages of various groups manifesting a particular attribute. For example, the proportion of *S*s answering "yes" to a questionnaire item in one psychiatric category might be compared with the proportion in another psychiatric category; or vocational groups might be so compared.

The traditional procedure has been to use the formula for the standard error of the difference between two observed proportions, although many authors of statistical texts (e.g., **2**, p. 54) add that the sampling distribution becomes highly skewed for extreme proportions, e.g., greater than .90 or less than .10. This makes use of tables of the normal distribution (or *t* distribution) quite inaccurate in these situations.

However there has been available for several years a transformation, known as the inverse sine function (**1**, p. 165), which tends to satisfy the condition of normality of experimental errors and thus makes the normal distribution tables useful for comparing even extreme proportions. Yet psychologists have evidently made little use of this function. In one of its forms, Zubin's *t* (**3**) (which is *not* the same as the Student-Fisher *t*), the function is defined as

$$t = 2\,sin^{-1}\sqrt{p},$$

in which p is the observed proportion and t is expressed in radians. The standard error of t is approximately $1/\sqrt{N}$, and for comparison of two proportions the following ratio is approximately normally distributed:

$$\frac{t_1 - t_2}{\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}}.$$

Since the sampling distribution of this function involves no parameters other than the number of observations, confidence limits can be set at once for an entire list of test items without the bother of calculating the standard error for each item separately as in the traditional procedure.

Table 1 enables the researcher to write down Zubin's *t* directly from the two frequencies (e.g., "yes" and "no") using only a slide rule. A certain amount of accuracy is sacrificed, but this error is small relative to the sampling error unless the number of observations is quite large. The table gives interval boundaries for *t* intervals of .05 (which correspond to *p* intervals of .025 in the middle range) in terms of the corresponding *ratio* of the two frequencies. In the traditional procedure, a preliminary step of adding the two frequencies for use as the denominator of the proportion was necessary before division could be undertaken.

Researchers who often deal with a large number of observations may wish to construct a similar table with finer intervals of *t*. Therefore the procedure by which Table 1 was constructed for intervals of .05 will be described. In the first column of the work sheet were written the following values of *t* in radians: 3.125,

TABLE 1
ZUBIN'S *t* FOR RATIO INTERVALS

Ratio	*t*	Ratio	*t*	Ratio	*t*	Ratio	*t*
293.1				4.672		5.223	
	3.00		1.55		2.25		.80
143.9		1.096		4.110		6.003	
	2.95		1.50		2.20		.75
84.47		1.211		3.632		6.949	
	2.90		1.45		2.15		.70
55.50		1.340		3.223		8.116	
	2.85		1.40		2.10		.65
39.32		1.483		2.868		9.582	
	2.80		1.35		2.05		.60
29.12		1.643		2.562		11.44	
	2.75		1.30		2.00		.55
22.42		1.822		2.297		13.86	
	2.70		1.25		1.95		.50
17.69		2.026		2.063		17.05	
	2.65		1.20		1.90		.45
14.31		2.255		1.855		21.47	
	2.60		1.15		1.85		.40
11.80		2.516		1.672		27.82	
	2.55		1.10		1.80		.35
9.870		2.814		1.509		37.17	
	2.50		1.05		1.75		.30
8.337		3.160		1.363		52.19	
	2.45		1.00		1.70		.25
7.130		3.558		1.233		78.36	
	2.40		.95		1.65		.20
6.153		4.023		1.114		130.6	
	2.35		.90		1.60		.15
5.345		4.571		1.008		255.4	
	2.30		.85		1.55		

3.075, 3.025, 2.975, . . . , 0.125, 0.075, and 0.025. The next column contained the corresponding values of *t*/2 in degrees, obtained by multiplying each value in the first column by 28.648. Then the sine of each of the values in the second column was recorded in the third column (**4**). The fourth and fifth columns contained sin^2 and $1-sin^2$, respectively. Then the ratios for Table 1 were obtained by division. For the left column of Table 1 each sin^2 was divided by the corresponding $1-sin^2$; for the right column, $1-sin^2$ by sin^2.

Table 1 has been constructed so that the larger frequency must always be used as the numerator of the ratio. The purpose here is to enable greater use of the left side of the slide rule, which is easier to read. When the positively scored category (e.g., "yes") has the larger frequency, the left side of Table 1 should be used; when the negatively scored category (e.g., "no") has the larger frequency, the right side should be used. A *t* value of .00 is equivalent to a *p* value of .00, and a *t* of 3.1416 (pi) is equivalent to a *p* of 1.00; however for ease of calculation it is more convenient to use 3.15 for the latter.

As an example consider the "yes" and "no" frequencies given in Table 2 for five questionnaire items. For Item 1 the ratio 1.53 falls between

TABLE 2

Fictitious Example of Five Questionnaire Items

Item	"Yes"	"No"	Ratio	*t*
1	52	34	1.53	1.80
2	28	58	2.07	1.20
3	79	7	11.3	2.55
4	43	43	1.00	1.55
5	12	74	6.17	.75

1.672 and 1.509 on the left side of Table 1. This interval is represented by a *t* of 1.80. Similarly the ratio 2.07 falls between 2.026 and 2.255 on the right side, giving a *t* of 1.20. Although a separate column of ratios has been included in Table 2 for expository purposes, in practice the researcher can write down the *t* value for each item directly by inspection of the slide rule and Table 1. The *t* values to be compared can be subtracted rather easily "in one's head" if they are efficiently arranged on the data sheet. Then the absolute differences which exceed a certain value, determined by use of the standard error formulas given above, can be noted.

After using this function for a while, the researcher will find that he begins to think directly in terms of *t*s; e.g., a *t* of 2.50 represents a very high proportion of positive replies, and a *t* of 1.00 represents a moderately low proportion of positive replies.

REFERENCES

1. Johnson, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.
2. McNemar, Q. *Psychological statistics.* (2nd Ed.) New York: Wiley, 1955.
3. Zubin, J. Note on a transformation function for proportions and percentages. *J. appl. Psychol.*, 1935, 19, 213–220.
4. *Tables of Sines and Cosines to Fifteen Decimal Places at Hundredths of a Degree.* U. S. Department of Commerce, National Bureau of Standards. Washington: U. S. Government Printing Office, 1949.

Received August 13, 1956.

Psychological Bulletin
Vol. 54, No. 3, 1957

AN IMPROVED METHOD FOR DERIVING EQUAL-DISCRIMINABILITY SCALES FROM RATINGS[1]

FRED ATTNEAVE
Air Force Personnel and Training Research Center
AND DAVIS J. CHAMBLISS
University of Wisconsin

A variety of closely similar methods (**1**, **2**, **3**, **5**) has been employed to scale ratings according to the principle of Thurstone's Law of Comparative Judgment (**4**), i.e., with variability of judgment as the unit of measurement. These methods suffer from at least one common defect: the unit contains a spurious or irrelevant component arising from differences in the constant rating tendencies of individuals. In a typical situation some raters will use all the scale categories whereas others will concentrate their judgments in two or three; some of the latter raters will rate at the low end of the scale, some at the high end, and some in the middle. As a result, the dispersion of ratings for any given stimulus is greater than it would be if all the raters used the scale in the same way.

It would be clearly desirable to equate all the raters with respect to such "constant error" tendencies, and use only the remaining "variable error" component as a yardstick of discriminability. This might be accomplished by transforming the ratings of each individual observer into *z*-scores, and then considering the distribution of such *z*-scores for each stimulus. To convert the ratings of each individual into percentiles or simply into ranks is computationally easier, however, and equally effective. (The fact that many tied ranks will necessarily occur in a given observer's ratings does not detract from the procedure, but in fact simplifies it.) Some distribution of ranks will then be associated with each stimulus, and an equal-discriminability scale may be constructed by applying to the rank scale a transformation which renders all such distributions as nearly as possible equal in dispersion.

The complete scaling procedure consists of the following operations:

1. Ratings are converted into ranks for each individual observer, considered independently of all the others. This conversion amounts to assigning a rank to each of the individual's rating-scale categories: e.g., if the rating scale has only seven categories, then seven particular ranks will describe all the individual's ratings, though what these ranks are will depend upon the way in which his ratings are distributed. The formula for the rank of a particular category k, for a particular observer, is as follows:

$$r_k = \sum_{j=1}^{j=k-1} n_j + .5n_k + .5,$$

in which n_k is the number of ratings placed in category k, and the $\sum n_j$ term is the total number of ratings placed in lower categories, by this one observer.

[1] This report is based on work done under ARDC Project No. 7706, Task No. 27001, in support of the research and development program of the Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

It may be noted, in passing, that the present step would be eliminated if the observers actually placed the stimuli in a complete rank order instead of using a rating scale, but that subsequent steps would be entirely applicable to data so obtained.

2. For each stimulus, a distribution of the ranks assigned to that stimulus by the various observers is plotted. Measures of central tendency and dispersion are found for each such distribution. ***Median*** and ***interquartile range*** are suggested as preferable to mean and standard deviation not only because they are easier to compute, but also because the rank distributions of high and low stimuli typically display a marked skewness toward the middle of the scale. Moreover, the distributions tend to be somewhat leptokurtic even when they are not skew. (Perhaps this is an overformal way of saying that there is usually some "lunatic fringe" of observers whose ratings fall "outside the distribution.") In view of this leptokurtic tendency, whatever its source, we have felt safer in using the interquartile range than a wider one, though in the case of a normal distribution the interval between the 7th and 93rd percentiles is the most stable.

3. A graph is prepared on which the reciprocal of the interquartile range of each rank distribution is plotted against the median of the distribution, as in Fig. 1. The reason for using the reciprocal is that we wish to deal with number of interquartile range units per rank, rather than with number of ranks per interquartile range unit. The points may be expected to fall into a **U**-shaped function which indicates that extreme stimuli are judged more precisely than those of intermediate value. A smooth freehand curve is drawn to fit these points; in the empirical cases with which we have dealt, the data have determined such a curve with relatively little ambiguity (see Fig. 1). The fitting process may be facilitated by the use of either a "rolling" average or of averages over certain fixed intervals.

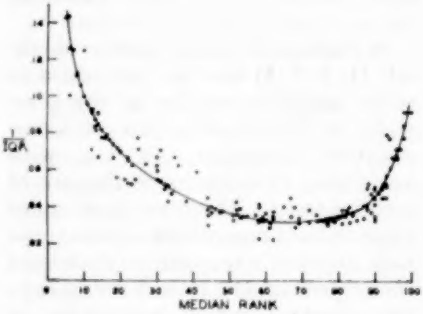

FIG. 1. RECIPROCAL OF INTERQUARTILE RANGE OF STIMULUS RANK DISTRIBUTION AS A FUNCTION OF THE MEDIAN OF THE DISTRIBUTION

4. Next the integral or cumulative area of this curve above some arbitrary origin is obtained. A table giving the cumulative area of the curve for each rank may be compiled with no great effort by adding successive ordinates of the curve, cumulatively, on a desk calculator. The values so obtained will be approximate, but highly accurate.

Suppose, for example, that the lowest stimulus has a median rank of 8. We may start integrating at this point, and assign rank 8 a cumulative area value of zero. The value of the integral for rank 9 will be equal to the ordinate (read from the graph) at 8.5. The value for rank 10 will be equal to the ordinate at 8.5 plus the ordinate at 9.5, and so on.

5. This table enables us finally to transform the median ranks of the stimuli into values on an equal-discriminability scale. Each of the slices added in the integration process represents the (fractional) number of

interquartile range units between one rank and the next. Therefore the sum of these slices, up to a given rank, represents the number of interquartile range units between a stimulus with that median rank and the origin (which may, as suggested above, be placed at the lowest stimulus).

Interquartile range values may be transformed into equivalent standard deviation values simply by multiplying through by the constant 1.349, which is the ratio of these two measures for the normal distribution. For most purposes the values might as well be left in terms of interquartile range. In Fig. 2 the cumulative area under the function in Fig. 1 is graphed, with scale values expressed in both units on the ordinate.

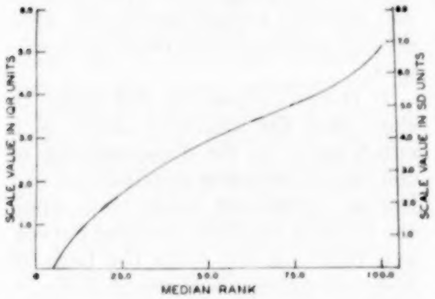

FIG. 2. SCALE VALUES CORRESPONDING TO MEDIAN RANKS OF STIMULI

In an experiment which will be reported elsewhere, 114 pairs of polygons were rated on a similarity-difference scale by 140 observers. Each pair was treated as a "stimulus," for scaling purposes. The ratings were scaled by both the graded-dichotomies method (1) and the present ranking method. (Figs. 1 and 2 are based on data from this experiment.) The whole rating and scaling procedure was then replicated with new observers. In the original experiment, scale values obtained for the stimuli by the graded-dichotomies method covered a range of 5.5 sigma units, whereas the ranking method yielded a range of 6.7 sigma units. In the replication, the graded-dichotomies method gave a 4.9 sigma spread, and the ranking method a 5.7 sigma spread.

The correlation between original and replication was .988 for the graded-dichotomies values, and .992 for values obtained by the ranking method. These figures imply that the amount of experimental-and-scaling error in the graded-dichotomies values was about 50% greater than that in the ranking method values (presumably most of the error was associated with random but genuine differences of judgment between the two samples of observers, rather than with scaling technique). It was further found that scale values obtained by the present method yielded higher correlations with certain physical measures on the stimulus material than did graded-dichotomies values. In short, the new method has been supported by all the empirical tests which have thus far been applied to it. Its primary justification should nevertheless remain the rational argument presented earlier.

REFERENCES

1. ATTNEAVE, F. A method of graded dichotomies for the scaling of judgments. *Psychol. Rev.*, 1949, **56**, 334–340.
2. GUILFORD, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1955.
3. SAFFIR, M. A. A comparative study of scales constructed by three psychological methods. *Psychometrika*, 1937, **2**, 179–198.
4. THURSTONE, L. L. The law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 273–286.
5. URBAN, F. M. The Weber-Fechner law and mental measurement. *J. exp. Psychol.*, 1933, **16**, 219–238.

Received for early publication March 4, 1957.

Psychological Bulletin
Vol. 54, No. 3, 1957

LEARNING CURVES—FACTS OR ARTIFACTS?[1]

HARRY P. BAHRICK
Ohio Wesleyan University

PAUL M. FITTS and GEORGE E. BRIGGS[2]
The Ohio State University

From an operational viewpoint, theory and method are intimately related. A theory has little value if no method is available for testing its predictions, and a method has no validity if it is not representative of the operations specified by some theory. Yet there are many instances in learning research in which behavior measures are chosen as a matter of convenience, rather than on theoretical grounds. This is done despite the fact that most theories of learning emphasize the distinction between the basic process of learning and indicants of this process (**12**, **16**, **17**) and despite the fact that the importance of this distinction is demonstrated in studies reporting low correlations among various indicants of presumably the same learning process (**5**; **10**, p. 138).

One of the most common instances of the arbitrary choice of response measures in learning studies is the use of a dichotomous score as an indicant of an underlying process which is known or assumed to be continuously distributed. We have reference to the use of arbitrary criteria of success and failure, or arbitrary criteria of the occurrence of a response, such as the extent of entry into a cul-de-sac necessary for recording an error in maze learning, the magnitude and latency of a reaction necessary for recording the occurrence of a conditioned response, or the size of the target used in determining the number of hits in a skilled task. A continuity viewpoint holds that the processes underlying these phenomena will produce a continuous and often normal distribution of response measures.

It is our purpose in this paper to show that the arbitrary choice of a cutoff point in the dichotomizing of continuous response distributions can impose significant constraints upon the shape of resulting learning curves, and that this can form the basis of misleading theoretical interpretations. We have chosen for illustration of this point the use of time-on-target scores as indicants of the level of skill attained in tracking tasks. However, we believe that the principles developed are quite general and apply to many learning situations.

Time-on-target scores reflect the amount of time during a trial that *S* is able to remain within an arbitrarily specified region around a target. A great many reports have been published during the last few years in which *E* has made use of such scores. In a few studies the effects of target size upon transfer of training have been examined (**3**, **8**). Two studies

[1] This research was conducted in the Laboratory of Aviation Psychology of The Ohio State University and was supported in part by the United States Air Force under Contract No. AF 41(657)-70 with the OSU Research Foundation, monitored by Dr. Robert L. French, who served as Scientific Officer for the AF Personnel and Training Research Center. Permission is granted for reproduction, publication, use, and disposal in whole or in part by or for the United States Government.

[2] The authors are indebted to Dr. David Bakan for many constructive criticisms.

(7, **14**) have dealt with the validity of time-on-target scores, and one of these (7) concluded that their usefulness is limited because the correlations between such scores and average error scores varies as a function of target size and problem difficulty. However, emphasis has not been placed on the importance of the constraints which the choice of a scoring zone exercises upon the shape of learning curves plotted from the recorded data and upon the conclusions which can be drawn from these functions.

It is our purpose here to show that the same tracking behavior, when scored with different target-tolerance standards, will result in learning curves which differ greatly in shape, and that the differing shape of learning curves obtained with various-sized scoring zones can be predicted theoretically from assumptions regarding the error-amplitude distributions. In further support of this view we present empirical data which indicate, in the case of tracking behavior, that the underlying error distribution to which all conventional scores can be related is continuously and normally distributed. Finally, we point out that a lack of understanding of these differential characteristics of response measures can easily lead to incorrect conclusions regarding the effects of other variables.

Empirical Data

The data reported here are taken from two studies (**1**, **5**) in which *S*s practiced one-dimensional tracking tasks on an electronic tracking apparatus described elsewhere (**18**). The tracking problems in the two studies varied in difficulty. We shall present first the learning curves obtained for the more difficult problem (**5**). In this study the target motion consisted of a 10 cpm sinusoidal motion of a line on a cathode-ray-tube (CRT) display. A filter with a time constant of .4 sec. introduced an exponential lag between the output of *S*'s arm control and its effect on tracking error. A compensatory display was used which provided a target line that remained stationary in the center of the display, and a cursor that moved to the right or left depending on the direction of the error from moment to moment. Two types of performance measures were taken on even-numbered 90-sec. trials: RMS error scores and time-on-target scores.

An electronic circuit provided a means of continuously obtaining the magnitude of the error (in the form of an electrical voltage), squaring this voltage, and integrating it over the period of a trial. The output of this circuit appeared on a voltmeter and the square root of this meter reading provided an index of the root mean squared error (RMS). The error voltage is computed with respect to an absolute reference of zero volts, whereas *S*'s error amplitude distribution may show some constant bias toward plus or minus voltages (i.e., error to the right or left of the target). As a result, the error RMS reflects both the variability of *S*'s distribution of amplitudes and any small constant error in average cursor position.

Time-on-target measures give the total time that the absolute magnitude of the error voltage was smaller than a given magnitude. Three such scores were taken for target zones of 5%, 15%, and 30% of ±5 v., which was the maximum problem voltage. These zones correspond to errors of .1, .3, and .6 in. of displacement of the cursor to either side of the target line, respectively. The three zones (from smallest to largest) will be referred to hereafter as zones A, B, and

C in order to avoid confusion when it is desired later to refer to the percentage of time *S* was "on target." Similarly, the time-on-target scores will be referred to as scores A, B, and C.

Fifty male and 50 female *S*s were used, and since the male *S*s were superior trackers on the average, separate learning curves are presented for the two groups. These curves are the empirical curves shown in Fig. 1 and Fig. 2 for the RMS scores and the three time-on-target scores.

The various learning curves suggest different accounts of the relative and absolute improvement during practice. The curves for time-on-target scores all suggest that absolute as well as relative improvements during tracking were greater for the male than for the female *S*s. This effect is particularly apparent for the smallest target zone (zone A) and becomes progressively less pronounced for the larger target zones. Between trials 2 and 14 the males improved by 33.2%, 31.9%, and 18.7% for scores in zones A, B, and C, respectively,

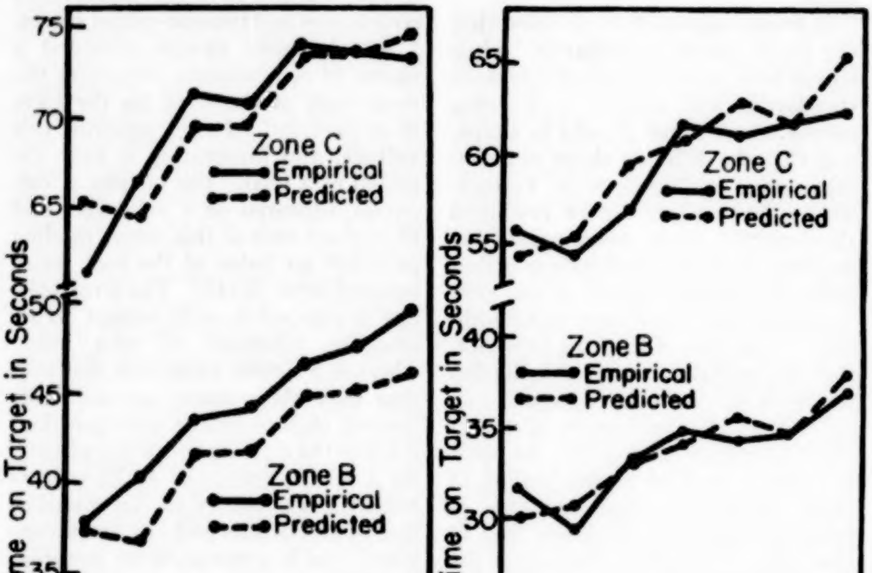

FIG. 1. TIME-ON-TARGET SCORES FOR 50 MALE *S*s (LEFT) AND 50 FEMALE *S*s (RIGHT) ON A DIFFICULT TRACKING TASK

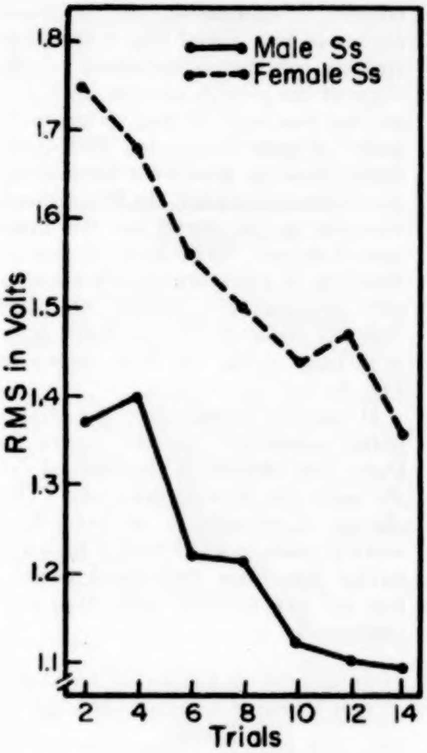

Fig. 2. RMS Scores on a Difficult Tracking Problem

while the corresponding improvements for the females are only 2.5%, 17.6%, and 11.8%. The RMS curves, however, indicate a greater improvement for the females, with a 22.3% reduction of error as contrasted with a 20.4% error reduction for the males. And all of these scores, it should be remembered, are derived from a single error voltage!

The widely divergent picture of the amount of improvement resulting from practice, given by the four scores described above, can be accounted for on theoretical grounds, which will be developed in the next section. Briefly, it will be shown that time-on-target scores are nonlinear, being relatively insensitive to changes in level of performance both above and below a critical region. Most of the female *S*s who served in the study referred to above, for example, were relatively poor trackers at the beginning of practice, and the zone A time-on-target score was almost completely insensitive to any improvement at this level of skill. Improvements at this level, however, were reflected in fairly large reductions in the error RMS score.

We shall now present data from the second study (1) which illustrate the same kind of scoring artifact with a less difficult tracking task. As in the first study, the problem was that of compensating for a 10 cpm target oscillation. However, no lag was introduced between the control output and the cursor movement. Several control loading conditions were used in this study as independent variables. We have selected four learning curves from the condition in which both a spring and a mass were used to load the control, since this condition of the study generally resulted in the best performance. Twenty-five males served as *S*s. The mean learning curves obtained for the RMS score and for three time-on-target scores employing the same relative target zones as were used in the previous study are shown as the empirical curves in Fig. 3.

It can be seen that the empirical curves in Fig. 3 again give different accounts of the improvement in performance at different stages of practice. The zone C curve is negatively accelerated and shows most of the improvement during the early trials, with smaller improvement during the last few trials. The curve for zone A, on the other hand, shows the largest gain during the last two trials, and relatively less gain during the early

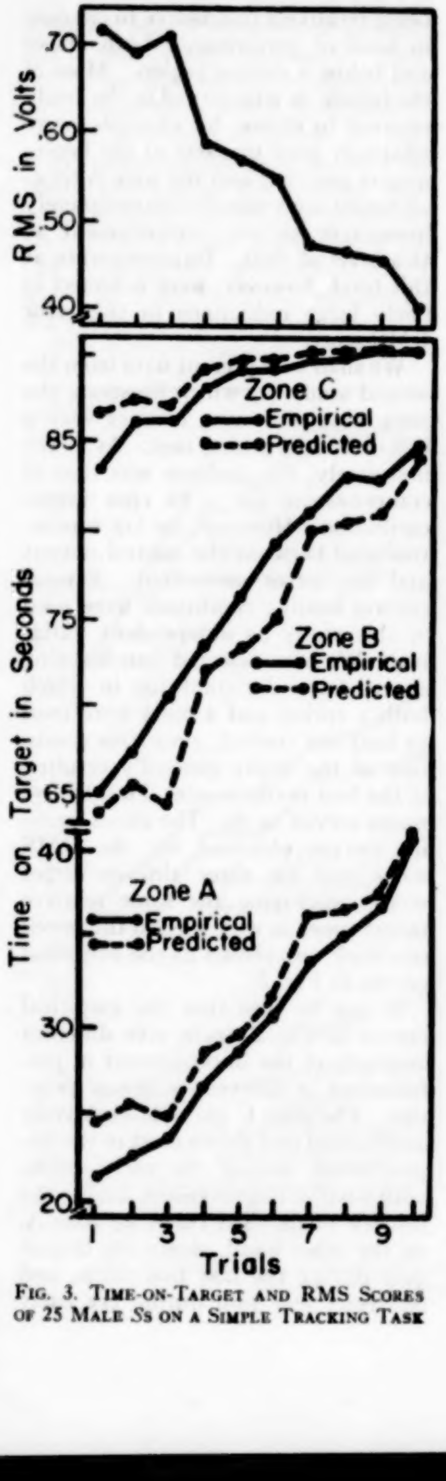

FIG. 3. TIME-ON-TARGET AND RMS SCORES OF 25 MALE *Ss* ON A SIMPLE TRACKING TASK

trials. In comparing the empirical curves of Fig. 1 and Fig. 3 the most striking contrast is furnished by the slope of the zone A curves. In Fig. 3 we can note a great deal of improvement of zone A scores, and these scores seem to provide a very sensitive index of learning. In Fig. 1 much less gain is registered for the same zone A scores. Particularly for female *Ss* in Fig. 1, this score reveals scarcely any improvement, despite the fact that the reduction of the RMS error is as large as for the data shown in Fig. 2.

It can be shown that the differential sensitivity of the scores in these two studies is determined by the variation in task difficulty. The change of sensitivity of individual scoring zones as a function of task difficulty has been mentioned earlier, but will now be dealt with more systematically.

PREDICTION OF LEARNING CURVES FOR VARIOUS TIME-ON-TARGET ZONES

If we assume that the amplitudes of tracking errors form a normal distribution during a trial, it is apparent that the percentage of this distribution which would fall within a given target zone can be determined, provided the standard deviation of the distribution of tracking errors is known. To illustrate the differential sensitivity of various scoring zones, we show in Fig. 4 predicted time-on-target scores for five target zones of differing size as a function of the magnitude of the RMS value of the error distribution.

Successive values for these curves are found by determining the ratio of the scoring zone, in volts, to the RMS values of the error distribution, also in volts. The ratios are *z* scores and the percentage of a normal dis-

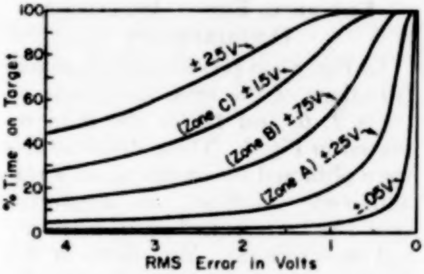

Fig. 4. Percentage of Time on Target for Various Target Zones as a Function of RMS of a Normal Distribution of Error Amplitudes

tribution between zero and each *z* score is found from a table of the normal curve. These values are multiplied by two to include errors on both sides of the target, and are plotted on the ordinate opposite the assumed RMS value.

It can be seen that each of the curves in Fig. 4 shows a maximal slope at a different range of variation of the RMS value, and becomes insensitive to variations outside that range. The ranges of maximal sensitivity shift toward smaller RMS values as we move from larger to smaller target zones. The sensitivity of a time-on-target score is maximal when the zone is of a size that includes ±1 *SD* of the error distribution, so that *S* is on target about 68% of the time. For smaller or larger target zones the score becomes progressively less sensitive to changes in the RMS value of the error distribution.

Functions similar to those shown in Fig. 4 can be plotted for target zones of any desired size, and it is apparent that curves for very small target zones would show their maximal sensitivity in an RMS range in which the curves of larger target zones have already approached an asymptote.

It is obvious that a score cannot reveal improvements once performance is approaching an asymptote of 100% time on target. However, the relative lack of sensitivity of each score at low performance levels is not generally recognized.

Empirical learning curves can be expected to depart from the curves shown in Fig. 4 for at least two reasons: (*a*) the theoretical curves in Fig. 4 are plotted for linear decreases in error RMS, while the observed decreases of error RMS during practice are usually a negatively accelerating function; and (*b*) the theoretical curves are also based upon the assumption that the amplitude distribution of error is normally distributed on all trials. In order to assess the significance of departures from normality in the data reported in Fig. 1 and Fig. 3, we have plotted curves for all three target zones using the observed RMS values for each trial and the corresponding ordinate values from the theoretical curves in Fig. 4. The divergence of the predicted curves from the corresponding empirical ones can be attributed in large measure to departures of the error distributions from normality. Unreliability of the electronic scoring equipment would, of course, also contribute to such divergence, but is believed to be quite small in the present case.

It can be seen that the empirical curves in Fig. 3 correspond moderately well to those predicted from the assumption of a normal distribution of error amplitudes. In Fig. 1 the correspondence of empirical and predicted curves is close in the case of the zone C curves. For the zones A and B curves, male *S*s performed better than would be predicted on the assumption of normality, and for the zone A curves this divergence is quite pronounced, particularly during the last few trials of practice. On the basis of our analysis we would expect these relations only if the amplitude distribution of tracking error were more peaked than a normal distribution, i.e., if the area near the center of the distribution were greater than predicted from the *z* scores. In order to check this prediction we shall need to examine in detail the empirical error-amplitude distributions of *S*s during tracking. This we proceed to do in the next section.

Empirical Error-Amplitude Distributions

In Fig. 5 we present empirical distributions of the error amplitude on trials 2, 6, and 14 for the data reported in Fig. 1. These distributions were obtained by means of an error-amplitude analyzer, an apparatus that has been described in more detail elsewhere (**5**). Also shown in this figure are normal curves with the same mean and *SD* as those of the

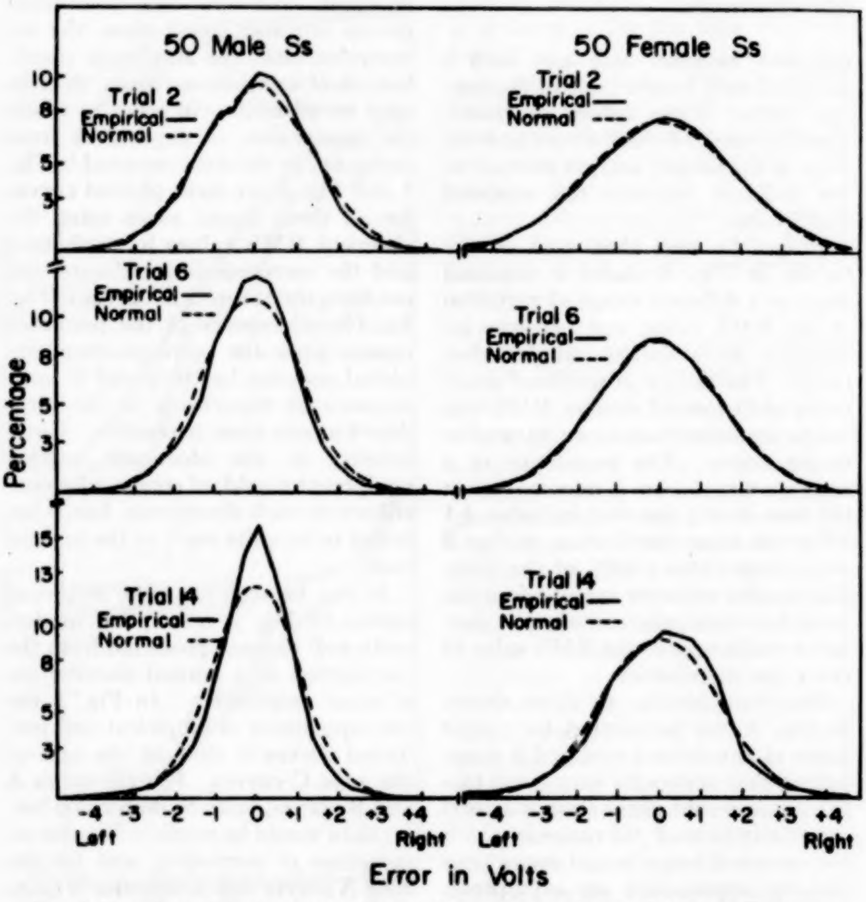

Fig. 5. Empirical Distribution of Error Amplitudes at Three Stages of Training

empirical distributions. Inspection of Fig. 5 confirms our theoretical prediction from the data of Fig. 4. It can be seen that the error distributions of female *S*s do not depart greatly from normality, and neither do the predicted values of their learning curves (Fig. 1) vary greatly from the observed ones. For male *S*s, however, the obtained distributions are more peaked than the corresponding normal ones, particularly on later trials. This confirms our interpretation of the departure of the empirical learning curve from the predicted curve for zone A scores (in Fig. 1).

Because the data of Fig. 5 are pooled for 50 *S*s, the peaking of the combined error-amplitude distribution may be the result of at least two different conditions. It is possible that this type of departure from normality characterizes the individual error distributions of the majority of *S*s, or it may be that all or most individuals show normal error distributions, but that we have a nonnormal distribution of individual differences among our 50 male *S*s. In other words, the combining of 50 normal distributions with different *SD*s can yield a curve such as we obtained, provided sizable proportions of these curves (i.e., individuals) represent unusually good and unusually poor performance.

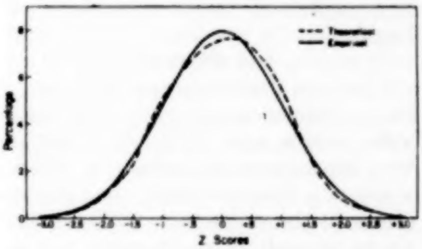

FIG. 6. ERROR AMPLITUDE DISTRIBUTION FOR MALE *S*s ON TRIAL 14, AFTER CONVERSION INTO *Z* SCORES

To determine which of these two situations prevailed we analyzed in more detail the data for the 50 males on trial 14, since this distribution (Fig. 5) shows the most pronounced departure from normality. The error amplitude distributions of each of the individual *S*s on this trial were converted into *z* scores after the *SD* of each *S*'s own amplitude distribution was determined. The ordinates for successive *z*-score values of .1 were averaged for the 50 *S*s and the resulting distribution is shown in Fig. 6, together with a normal distribution. It can be seen that the peaking effect is not due to departures from normality in the error-amplitude distributions of individual *S*s, but rather that it is due to the combining of normal distributions which among themselves are not normally distributed. We are apparently dealing with a situation in which individual differences are normally distributed early, but not later in learning.

The problems involved in interpreting learning curves based on group data have been the concern of several recent papers (**2**, **4**, **15**). In the present instance, however, we are chiefly interested in accounting for the departures of the obtained curves from the predicted curves in Fig. 1. The progressive peakings of the group amplitude distributions for male *S*s appear to be a sufficient explanation for this phenomenon. During the later stages of practice the change in the shape of the pooled amplitude distribution has made the zone A curve more sensitive and the zone C curve less sensitive than would be the case if the group error-amplitude distribution had remained normal.

Confounding of Effects Produced by the Manipulation of Independent Variables and Artifacts Produced by Scoring

We have shown in the preceding section that learning curves based upon a particular-sized time-on-target zone will be maximally sensitive over a relatively narrow range of RMS values, and be relatively insensitive to variations of RMS outside that range. This means that large differences in the RMS value of the error-amplitude distribution may exist and may result in small or large differences of performance on a time-on-target score, depending upon the sensitivity of the score over the RMS range in question. If we now use a time-on-target score as a means of determining the functional relation between two variables, functions at various stages of learning, or functions for several versions of a task which differ in difficulty, the change in sensitivity of our measure must be taken into account. Particularly must we guard against scoring artifacts when we look for interaction effects. Otherwise we may conclude that the independent variable produces important effects at one stage of training and not at other stages, or important effects on a simple-task version and not on a difficult-task version, when in fact we are dealing with artifacts produced by the nonlinearity of our measures.

There are many instances in the literature where authors have failed to consider the above effects in their interpretation of results based on time-on-target scores. In order to call attention to this problem we have chosen two reports of work done in our own laboratory.

In a recent paper (9) which evaluates the effect of stimulus and response amplitudes upon tracking performance, a 5% time-on-target score, covering an error voltage range of ±.325 v., was used. In discussing the interactions of stimulus and response amplitude upon performance, the authors conclude from statistical tests of their scores that as the stimulus amplitude was magnified, *S*s found it increasingly advantageous to use a large response motion. This conclusion appears quite reasonable if we examine in their Fig. 4 (9, p. 86) the progressive separation among response-amplitude curves for increasing values of stimulus amplitude. However, the small stimulus amplitudes resulted in scores of only 20% to 30% time on target, while time-on-target scores as high as 50% to 60% were achieved under conditions of large stimulus amplitude. If we now refer to Fig. 4 of the present article it can be seen that the acquisition curve for a time-on-target zone that provides scores near 50% on target is very sensitive to variations of RMS, whereas a time-on-target zone giving scores in the range near 20% is relatively insensitive to identical variations in RMS error. Thus, assuming a normal distribution of error amplitudes, slight variations in the RMS value of the error would produce large effects on the time-on-target score if the stimulus amplitude is large, but only small effects when the stimulus amplitude is small. The range of RMS values occurring in this study varied from about .4 to 1 v. If a larger target zone had been used, for example, one covering ±.75 v., it is possible that the statistical analyses would again have shown a significant interaction, but the obtained curves would have shown more separation for small than for large stimulus amplitudes and the authors would have been forced to make an opposite conclusion regarding the direction of the interaction effect among stimulus and response amplitudes.

In this same study the absence of significant stimulus- and response-amplitude effects upon performance in the compensatory version of the task may have arisen because of a similar artifact of scoring. It can be seen in Fig. 4 (**9**, p. 86) that time-on-target scores under the compensatory 30-plus-20-plus-10-cpm-frequency condition did not greatly exceed 10%. Inspection of Fig. **4** in the present article shows that the curve for a target zone giving 10% time-on-target scores has extremely poor sensitivity in terms of the RMS criterion. Using their particular target zone for this version of their task, it would be difficult to demonstrate the effect of any independent variable upon performance. We do not believe, therefore, that a comparison of the relative effect of amplitude factors on compensatory vs. pursuit versions of the tracking task is possible on this basis. Thus, whereas the authors were careful to use the same criterion measure (a particular time-on-target zone) for all of their task variations, the very use of a standard measure, which was differentially sensitive to tasks of varying difficulty, limits the validity of some of their conclusions. This should, however, not be interpreted as a criticism of the major findings of the study which do not depend upon assumptions of linearity.

In another study (**12**) evaluating the effects of control loading upon tracking performance, a similar effect can be observed. In summarizing their results the authors conclude that the differential effects of control loading upon performance seem to increase during the first 20 practice sessions. We have reproduced their Fig. 2 (**12**, p. 355) as Fig. **7** of the present report. Inspection of Fig. **7** indicates that the various curves show increasing separation as practice progresses, and this fact forms the basis for the above conclusion of the authors. However, the initial performance level is only about 12% time on target, a range in which time-on-target scores are insensitive, while the terminal performance level is one which brings the scores into a much more sensitive range of the performance measure. To illustrate that this increasing separation among learning curves is an artifact of the gradually increasing sensitivity of the scoring zone, the curves have been replotted in Fig. **8** for a larger scoring zone.

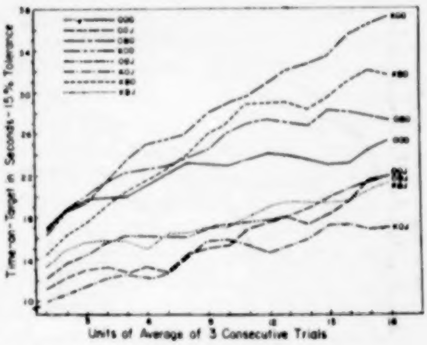

FIG. 7. MEAN 15% TOLERANCE SCORES OF NOVICE *S*s FOR EIGHT EXPERIMENTAL CONDITIONS FROM HOWLAND AND NOBLE (**12**)

The conversion of scores is based upon the functions shown in Fig. **4**, and involves the assumption of a normal distribution of error amplitudes. It can be seen that the increasing separation among curves is no longer present in Fig. 8. Thus the authors have attributed an artifact produced by the arbitrary selection of target size to their independent variable, and they might have reached an opposite conclusion had a different target zone been selected.

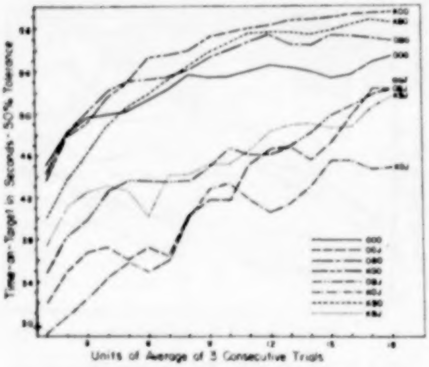

Fig. 8. The Data of Fig. 7 Replotted for a 50% Target Tolerance

Implications for Performance Measurement

It should be pointed out that the nonlinear relation between RMS and time-on-target scores does not invalidate all use of the latter scores. For certain gross comparisons, intended only to determine the presence or absence of a significant effect, either type of score may be adequate. Indeed, the two types of scores would be expected, and have been found (**5**, **6**, **7**) to correlate rather highly. Artifacts in the interpretation of results occur primarily when attempts are made to test for interaction effects or to interpret functional relations over an extended range of task difficulty or over an extended period of learning.

Thus it would appear that a single target zone can provide a score of only limited value as an indicant of tracking performance. This is particularly true if performance on different tasks or at different stages of learning varies over a wide range, so that the percentage of time on target is either very low, or very high for some of the conditions to be evaluated.

Simultaneous recording of scores for several target widths is one method of obtaining performance records which are less likely to lead to confounding of the effects produced by independent variables and the scoring mechanics (although perhaps this presents *E* with a difficult choice of functions). Another possible approach would involve transformation of time-on-target scores to yield an estimate of the RMS score, by constructing a plot such as that shown in Fig. 3, or by direct reference to a table of the normal curve. This procedure may be of limited value, however, because of excessive demands upon the reliability of the scoring apparatus near the extremes of the scale.

The use of the RMS measure itself appears to us as the best method of avoiding the problems discussed in this paper. The best single statistic (in addition to the mean) for describing a normal or near-normal distribution is generally accepted to be the *SD*. Furthermore, this score does not impose an artificial ceiling upon improvement as do the time-on-target scores. Other advantages lie in the fact that the RMS value provides a score equally useful for problems of all difficulty levels, and that the measure reflects the entire distribution of error amplitudes, rather than just a dichotomized version of the distribution. The selection of this measure is, of course, also arbitrary in one sense, since RMS does not change linearly as a function of practice. The lack of true scales for the measurement of learning, and the consequent difficulty of comparing variability at different stages of practice has been pointed out before (**11**, p. 635), and the use of the RMS measure does not solve these problems. The advantage of the RMS measure simply lies in the substitution of a single function for an unlimited number of functions determined by all possible target dimensions. As a consequence, there result advantages of comparability of data and ease of interpretation. If the RMS score is computed with respect to zero error equals perfect performance, rather than with respect to zero equals *S*'s own mean, then the score will reflect constant error as well as variable error (i.e., $MS_{\text{total error}} = MS_{\text{variable error}} + MS_{\text{constant error}}$). It would appear from the amplitude distributions presented here that such constant errors are relatively minor and can usually be disregarded. This is likely to be true in most studies of continuous tracking, where lead or lag of the cursor relative to the target will each result in positive and negative voltage errors depending upon the momentary direction of motion. However, the mean plus or minus error can usually be determined by the use of slightly more complex scoring equipment and the RMS score can then be reduced to the variable error in pure form. Thus, we can conclude that the best single measure of tracking performance is error RMS (or perhaps simply mean error). A more complete picture of performance can be gained by recording the complete amplitude distribution of error.

Although the present analysis has dealt only with the measurement of tracking performance, the conclusions have much wider implications for psychology. Similar problems exist wherever response characteristics follow a continuous and normal distribution and where learning results in diminished variance of this distribution, but performance is scored according to an all-or-none criterion of frequency of occurrence. Not only are performance scores in tracking tasks, such as that provided by the rotary pursuit apparatus, subject to artifacts arising from the arbitrary choice of the size of the target zone, but so are scores in many other tasks such as steadiness tests, dotting tests, tweezer dexterity tests, pegboard tests, etc., where success is scored against an all-or-none criterion. It even appears likely that the records of many other types of behavior, including such diverse responses as conditioned eyelid responses, leg flexion, and maze turning, which are recorded in terms of frequency of occurrence, may show similar artifacts provided the underlying habit strength varies as a continuous function of practice.

REFERENCES

1. Anderson, Nancy. Factors of motor skill learning related to control loading. Doctor's Dissertation, The Ohio State University, 1956.
2. Bakan, D. A generalization of Sidman's results on group and individual functions and a criterion. *Psychol. Bull.*, 1954, **51**, 63–64.
3. Bilodeau, E. A. Accuracy of response as a function of target width. *J. exp. Psychol.*, 1954, **47**, 201–208.
4. Estes, W. K. The problem of inferences from curves based on group data. *Psychol. Bull.*, 1956, **53**, 134–141.
5. Fitts, P. M., Bennett, W. F., & Bahrick, H. P. Application of autocorrelation and crosscorrelation analysis to the study of tracking behavior. In G. Finch & F. Cameron (Eds.). ***Symposium on Air Force human engineering, personnel, and training research.*** Baltimore: Air Research and Development Command, 1956. Tech. Tep. 56–8.
6. Fitts, P. M., Marlowe, E., & Noble, M. E. The interrelations of task variables in continuous pursuit tasks: I. Visual-display scale, arm-control scale, and target frequency in pursuit tracking. *USAF, Hum. Resour. Res. Cent., Res. Bull.* Sept., 1953, No. 53-34.

7. Gray, Florence E., & Ellson, D. G. The validity of time-on-target (clock) scores as an estimate of tracking error magnitude. USAF, Air Materiel Command, Wright-Patterson AFB, *Memo. Rep.* TSEAA-694-2F, June, 1947.

8. Green, R. F. Transfer of skill on a following tracking task as a function of task difficulty (target size). *J. Psychol.*, 1955, **39**, 355–370.

9. Hartman, B. O., & Fitts, P. M. Relations of stimulus and response amplitude to tracking performance. *J. exp. Psychol.*, 1955, **49**, 82–92.

10. Hilgard, E. R., & Marquis, D. G. *Conditioning and learning.* New York: Appleton-Century, 1940.

11. Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Ch. 17.

12. Howland, D., & Noble, M. E. The effect of physical constants of a control on tracking performance. *J. exp. Psychol.*, 1953, **46**, 353–360.

13. Hull, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.

14. Humphrey, C. E., Thompson, J. E., Ensor, H. L., & Versace, J. The measurement of tracking error: time-on-target. Johns Hopkins Univer., *Applied Physics Lab. Rep.* APL/JHU TG-196, April, 1953.

15. Sidman, M. A note on functional relations obtained from group data. *Psychol. Bull.*, 1952, **49**, 263–269.

16. Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York, 1951. Ch. 1.

17. Tolman, E. C. The determiners of behavior at a choice point. *Psychol. Rev.*, 1938, **45**, 1–41.

18. Warren, C. E., Fontaine, A. B., & Clark, J. R. A two-dimensional electronic pursuit apparatus. *USAF, Hum. Resour. Res. Cent. Res. Bull.* August, 1952, No. 52-26.

Received for early publication March 5, 1957